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

На правах рукописи

Карасев Дмитрий Алексеевич

Разработка метода протеохемометрики для предсказания взаимодействий белков и лигандов на основе их локального сходства

1.5.8. – Математическая биология, биоинформатика

Диссертация

на соискание ученой степени кандидата биологических наук

Научный руководитель:

доктор биологических наук, профессор РАН

Лагунин Алексей Александрович

МОСКВА – 2023

ОГЛАВЛЕНИЕ

СПИСОК СОКРАЩЕНИЙ И УСЛОВНЫХ ОБОЗНАЧЕНИЙ

БД – База данных

МО – машинное обучение

(Q)SAR – Quantitative structure–activity relationship ((количественное) моделирование структура-активность)

GPCR – G protein-coupled receptors (рецепторы, сопряжённые с G-белком)

IC50 – концентрация полумаксимального ингибирования

Kd – константа диссоциации

Ki – константа ингибирования

MNA – Multilevel Neighborhoods of Atoms

PASS – Prediction Activity Spectra of Substances

PCM – Proteochemometrics (протеохемометрика)

SPrOS – Specificity Projection On Sequence

ВВЕДЕНИЕ

Актуальность и степень разработанности темы исследования

Компьютерные методы играют важную роль при поиске биологически активных химических соединений и идентификации их молекулярных мишеней [Sun и соавт., 2022]. Такие методы применяются на различных этапах разработки лекарственных соединений. Например, при виртуальном скрининге больших массивов химических структур, для исследования производных какого-либо базового соединения [Sliwoski и соавт., 2013]. В последние годы набирает популярность репозиционирование уже известных лекарственных соединений для новых мишеней [Oprea и соавт., 2012, Savosina и соавт., 2021].

Методы, в которых используются трехмерные структуры белков мишеней и лигандов, предполагают исследование расшифрованного комплекса, либо моделирование взаимодействия лиганда с новой для него мишенью. Это позволяет получить большое количество полезной информации, включая локализацию лиганда при связывании с мишенью, тип связывания [Wade и соавт., 2019]. Возможно, также моделировать взаимодействие мишеней с соединениями из виртуальных библиотек [Bender и соавт., 2021]. Существенным ограничением для таких подходов является отсутствие разрешенной трехмерной структуры либо надежной модели исследуемого белка мишени. Для многих белков пока это не представляется возможным [Duran-Frigola и соавт., 2013].

Для массовых предсказаний взаимодействий белок-лиганд используются обучающие данные, составленные из химических структур лигандов, классифицированных по названиям мишеней [Muratov и соавт., 2020]. Методы, основанные на структуре лиганда (моделирование структура-активность), предполагают наличие известных лигандов для исследуемой мишени, в противном случае построить модель не представляется возможным [Lapinsh и соавт., 2003, Bongers и соавт., 2019]. Чтобы преодолеть отмеченные ограничения исследователи разрабатывают методы, в которых наряду с информацией о структурах лигандов включаются сведения и о белках-мишенях, обычно об их

аминокислотных последовательностях [Westen и соавт., 2011]. Такой подход известен как «протеохемометрическое моделирование» (PCM) или «протеохемометрика» [Lapinsh и соавт., 2001]. Соответствующие алгоритмы нашли применение при анализе разнообразных данных по белок-лигандным взаимодействиям.

Существенной проблемой данной методологии является описание последовательностей белков [Bongers и соавт., 2019]. Исследователи по-разному подходят к решению этой задачи. Методика описания часто определяется количеством исследуемых последовательностей и их доменным составом. В случае с большим числом лигандов при относительно небольшом числе близко гомологичных белков-мишеней весьма эффективно использование множественного выравнивания. При этом выявляются позиции, консервативные в подгруппах лигандной специфичности [Nabu и соавт., 2014].

При увеличении количества последовательностей в выравнивании или при работе с эволюционно дивергировавшими белковыми семействами, например, протеинкиназами, не всегда удается получить приемлемое для анализа выравнивание. Большое число колонок (позиций), богатых разрывами создает информационный шум, который можно снизить, исключив такие колонки из дальнейших расчетов [Lapinsh и соавт., 2010].

При сильной дивергенции сопоставляемых последовательностей, выраванивание не всегда позволяет корректно совместить функционально важные позиции, особенно те, что специфичны для исследуемых подгрупп. В этом случае используются интегральные показатели, описывающие белок. К ним относятся аминокислотный состав, дипептидный состав, псевдоаминокислотный состав и др. [Tresadern и соавт., 2017]. Ряд исследователей использует ковариации и кроссковариации, основанные на физико-химических свойствах аминокислотных остатков [Lapinsh и соавт., 2010, Zakharov и соавт., 2019, Shaikh и соавт., 2016, Reker и соавт., 2017, Kim и соавт., 2020].

Применение интегральных оценок для описания белковой молекулы приводит к потере информации о вкладе отдельных аминокислотных остатков.

При этом хорошо известно, что единичные остатки вносят существенный вклад в аффинность низкомолекулярного лиганда к белку-мишени. Более того, отмечаются случаи влияния удаленных от области связывания аминокислотных остатков на аффинность [Karaman и соавт., 2008].

В настоящее время не разработано универсального метода, позволяющего работать с любыми группами белков-мишеней [Bongers и соавт., 2019]. Авторам приходится проводить предварительный анализ для оценки представительности исследуемых белков, их доменного состава, а также характеристик выравнивания. Это существенно ограничивает прогностические возможности существующих подходов, что, в свою очередь может сказаться на сроках разработки лекарственных средств, остро необходимых, например, при выявлении новых инфекционных агентов. Таким образом, насущной потребностью является разработка метода с широкой областью применимости, включающей группы мишеней и их лигнадов и разной степенью гетерогенности. Исходя из этого, мы сформулировали следующую цель исследования.

Цель и задачи исследования

Целью диссертационной работы является создание метода для широкомасштабного предсказания белок-лигандных взаимодействий на основе анализа локального сходства аминокислотных последовательностей белков и структур низкомолекулярных лигандов.

Для достижения цели исследования нами были поставлены и решены следующие **задачи**:

1. Сформировать обучающие выборки, содержащие информацию о структурах низкомолекулярных лигандов, аминокислотных последовательностях белков-мишеней и показателях аффинности для каждой пары «белок-лиганд».
2. Разработать метод для прогноза белок-лигандных взаимодействий на основе анализа локального сходства аминокислотных последовательностей белков-мишеней и структур низкомолекулярных лигандов.
3. Оценить эффективность разработанного метода при широкомасштабном прогнозе белок-лигандных взаимодействий на наборах данных, характеризующих взаимодействия лигандов с белками различных таксономических групп.
4. Реализовать веб-сервис для прогноза белок-лигандных взаимодействий на основе разработанного метода протеохемометрики.

Научная новизна

Разработан оригинальный метод протеохемометрики, который позволяет прогнозировать белок-лигандные взаимодействия для различных групп белков-мишеней. Методика прогноза основана на поиске локальных соответствий между атомами низкомолекулярных лигандов и аминокислотными остатками белков-мишеней. При этом не требуется модификации или оптимизации параметров при смене группы белков-мишеней, что является преимуществом в сравнении с существующими подходами. Эффективность нового метода продемонстрирована при тестировании наиболее типичных ситуаций, возникающих при компьютерной оценке взаимодействий белок-лиганд. Метод эффективно работает при прогнозировании спектра лигандов на основе аминокислотной

последовательности белка-мишени. Привлечение данных по структурному сходству лигандов, позволяет предсказывать новые пары белок-лиганд в отсутствии сведений о спектрах взаимодействия для обоих компонентов.

Впервые разработан свободно доступный в сети Интернет веб-сервис (http://way2drug.com/proteochemometrics/), который предоставляет пользователям широкий спектр возможностей для компьютерной оценки белок-лигандных взаимодействий на основе протеохемометрики.

Теоретическая и практическая значимость

Метод позволяет проводить фундаментальные исследования с целью изучения феномена белок-лигандных взаимодействий, оценивать вклад отдельных остатков в специфичность связывания и исследовать селективность различных ингибиторов к белкам-мишеням.

При создании новых лекарственных средств предложенный метод позволяет отбирать соединения, наиболее перспективные для экспериментального тестирования в отношении не только уже известных фармакологических мишеней, но и в отношении новых белков-мишеней, т.е. таких, для которых неизвестны низкомолекулярные лиганды. Прогноз возможен с использованием различных входных данных в зависимости от задач планируемого экспериментального исследования. Входной информацией являются либо аминокислотные последовательности, либо структуры химических соединений, либо данные обоих типов. Метод не требует оптимизации для новых групп белков, что позволит исследователям оперативно осуществлять прогноз для новых мишеней и отвечать на новые вызовы, связанные с поиском биологически активных соединений. Свободно доступный в сети Интернет веб-сервис предоставляет разработанный инструмент широкому кругу исследователей.

Личный вклад автора

Автор самостоятельно провел поиск и анализ литературы по проблемной области, провел обобщение современных достижений в области протеохемометрики и сформулировал пути решения существующих проблем. Автор сформировал программный комплекс для прогноза белок-лигнадных взаимодействий, разработал методику сбора данных. Все расчёты, построение моделей и анализ полученных результатов выполнены лично автором.

Методология и методы диссертационного исследования

Разработан программный комплекс для прогноза белок-лигандный взаимодействий, использующий оригинальные методы машинного обучения. Оценка эффективности предложенного выполнена с помощью процедуры скользящего контроля в соответствии с современными требованиями к валидации результов прогноза. Оценка области применимости предложенного похода осуществлена путем тестирования согласно трем наиболее типичным сценариям протеохемеметрики на нескольких наборах данных, характеризующихся разной степенью дивергенции последовательностей белков-мишеней. Для отбора тестовых и обучающих данных с целью их унификации и повышения надежности разработан и применен оригинальный метод.

Положения, выносимые на защиту

1. Разработан оригинальный протокол сбора наиболее надежных данных из доступных информационных источников и их унификации для создания классификационных протеохемометрических моделей.
2. Разработанная и реализованная в виде программного комплекса методика позволяет предсказывать белок-лигандные взаимодействия в соответствии со сценариями, использующими в качестве входных данных структуры лигандов и аминокислотные последовательности. Сценарий протеохемометрики, при котором используются входные данные обоих типов, рассчитан на наиболее частую ситуацию, связанную с неполнотой используемой при обучении информации.

3. Предложенный подход обеспечивает высокую точность предсказаний в широкой области применимости, которая включает наборы белков-мишеней с разной степенью структурно-функционального сходства. При моделировании ситуации с неполными обучающими данными показана высокая эффективность разработанного нами протеохемометрического подхода.

4. Свободно доступный в сети Интернет веб-сервис предоставляет широкому кругу исследователей возможность проведения протеохемометрического анализа (http://way2drug.com/proteochemometrics/).

Степень достоверности и апробация результатов

Точность разработанного подхода оценивалась с помощью скользящего контроля с исключением по-одному. Вклад аминокислотных остатков оценивался с помощью расчета *p*-уровня значимости.

По материалам диссертации опубликовано пять печатных работ. Основные положения диссертации были представлены на российских и международных конференциях и симпозиумах, включая: The 13th International Conference on Bionifomatics of Genome Regulation and Structure/Systems Biology (BGRS\SB-2022), Новосибирск (Россия), 2022; VII Съезд биохимиков России и X Российский симпозиум "Белки и пептиды", Дагомыс (Россия), 2021; Международный форум: Биотехнология: состояние и перспективы развития, Москва (Россия), 2020; 9-ая Московская конференция по вычислительной молекулярной биологии (MCCMB'19), Москва (Россия), 2019; 8-ая Московская конференция по вычислительной молекулярной биологии MCCMB'17, Москва (Россия), 2017; 43rd FEBS Congress, Biochemistry Forever, Прага (Чехия), 2018; VIII российский симпозиум «белки и пептиды», Москва (Россия), 2017; The 10th International Conference on Bioinformatics of Genome Regulation and Structure/Systems Biology (BGRS\SB-2016), Новосибирск (Россия), 2016; XXIII Российский национальный конгресс «Человек и лекарство», Москва (Россия), 2016.

Структура и объем диссертации

Диссертационная работа состоит из введения, обзора литературы, описания материалов и методов, результатов и обсуждения, заключения, выводов и списка цитированной литературы. Работа изложена на 101 странице, включает 23 рисунка и 7 таблиц. Список литературы содержит 155 литературных источников.

ГЛАВА 1. ОБЗОР ЛИТЕРАТУРЫ

1.1 Компьютерные методы, используемые при разработке лекарственных соединений

Результаты геномных, фармакогеномных, протеомных и метаболомных исследований являются неисчерпаемым источником информации для научного сообщества. На сегодняшний день собрано большое количество такого рода данных, которые поступают из самых разных источников от небольших лабораторий до крупных международных консорциумов [Richard и соавт., 2021, Hanash и соавт., 2004, Davis и соавт., 2011]. Исследования данной информации с применением статистических подходов позволяет персонализировать лечение пациентов [Curtis и соавт., 2012], предлагать новые методы лечения [Romond и соавт., 2005], а также создавать новые лекарственные соединения [Neves и соавт., 2018].

В последние десятилетия информационные технологии вместе с ростом доступных вычислительных мощностей создали основу для скрининга *in silico* больших химических библиотек. Компьютерные технологии с одной стороны увеличивают пространство для поиска новых препаратов, с другой стороны позволяют уменьшить количество соединений для доклинического исследования и тем самым позволяют существенно снизить финансовые затраты [Vamathevan и соавт., 2019]. В этом контексте методы машинного обучения (МО) получили большое распространение с целью планирования исследований по медицинской химии (рис. 1).

Рисунок 1. Использование методов машинного обучения при планировании экспериментов.

На сегодняшний день множество методов МО используются в фармакологических исследованиях для предсказания молекулярных характеристик, биологической активности, межлекарственных взаимодействий и побочных эффектов [Dara и соавт., 2022]. Наиболее распространенные алгоритмы МО используют наивный байесовский классификатор, метод опорных векторов, алгоритм случайного леса и нейронные сети [Blanco и соавт., 2018, Munteanu и соавт., 2010, Garcia и соавт., 2009, Liu и соавт., 2017, Riera-Fernández и соавт., 2011, Shirvani и соавт., 2020, Suay-Garcia и соавт., 2020].

Работа с МО включает в себя несколько этапов: сбор данных, расчет дескрипторов, построение прогностической модели, валидация модели. Для некоторых методов необходима процедура отбора признаков.

1.2 Анализ взаимосвязи структура-активность (Q)SAR. Границы применимости

Эффективным и широко применяемым подходом является предсказание биологической активности химического соединения на основе его структуры или определение взаимосвязи «структура-активность» [Muratov и соавт., 2020]. Более 60 лет назад, в 1962 г., Ханш и соавт. предложили метод, позволяющий предсказывать коэффициента распределения вода/октанол (logP) от структурных характеристик низкомолекулярного соединеия с помощью регрессионного анализа [Hansch и соавт., 1962]. Соответствующая публикация ознаменовала новый подход к компьютерной оценке свойств низкомолекулярных лигандов.

Оценка зависимости «структура-активность» (Structure-Activity Relationships, SAR) строится на двух базовых идеях: «структура молекулы определяет ее биологическую активность» и «структурно сходные молекулы обладают похожей биологической активностью». Различные модели позволяют получать качественную (SAR) или количественную (Quantitative SAR, QSAR) оценку в отношении физико-химических и биологических характеристик соединения на основе его химической структуры. Оба упомянутых подхода,

объединенные термином (Q)SAR, представляют собой очень широкий набор вычислительных инструментов, которые могут давать на выходе разнообразные прогнозируемые характеристики, получая входные данные в форме молекулярных дескрипторов. Модели (Q)SAR позволяют анализировать уже существующие молекулы, чтобы определить новые свойства, а также понять неочевидные взаимосвязи между их структурой и биологической активностью.

Для проведения исследования (Q)SAR необходимы три типа данных [Cronin и соавт., 2003]:

1. Структуры химических соединений;
2. Данные о биологической активности каждого из соединений;
3. Набор дескрипторов, для описания молекулярных структур.

Широкое распространение получил прогноз биологической активности еще не синтезированных соединений из виртуальных библиотек [Patel и соавт., 2020]. Последние могут генерироваться за короткое время, а применение к ним (Q)SAR моделей выполняется с условием, что структурные характеристики виртуальных соединений не должны выходить их за пределы применимости модели [Patel и соавт., 2020].

Применение (Q)SAR моделей сокращает расходы на лабораторное оборудование и реагенты. Значительная часть программного обеспечения для создания моделей (Q)SAR бесплатно предоставляется в сети Интернет. При этом, расчеты требуют незначительных временных затрат [Tetko и соавт., 2017]. В течение последних 50 лет ряд новых лекарств, принятых в медицинскую практику, был получен с применением методов (Q)SAR. Так, среди одобренных лекарств, разработанных с применением (Q)SAR, можно указать следующие препараты: ингибитор карбоангидразы – Дорзоламид (принят в 1995) [Vijayakrishnan и соавт., 2009, Ghosh и соавт., 2014]; ингибитор ангиотензинпревращающего фермента – Каптоприл (1981) [Talele и соавт., 2010]; три средства против вируса иммунодефицита человека (ВИЧ) – Саквинавир (1995 г.), Ритонавир и Индинавир (1996 г.) [Van Drie и соавт., 2007]; антиагренгант Тирофибан (1998 г.) [Hartman и соавт., 1992].

Одно из важнейших направлений (Q)SAR – предсказание взаимодействия низкомолекулярных соединений (лигандов) и биологических макромолекул (мишеней), среди которых чаще всего исследуются белки [Pogodin и соавт., 2019]. В этом случае рассчитывается прогностическая оценка специфичности тестируемого лиганда к мишеням обучающей выборки, идентфикаторы которых в данном случае выступают как класс-образующие признаки. При выполнении таких исследований возникают некоторые ограничения. Обучающие данные определяют взаимодействие конкретной мишени с группой лигандов без учета сходства между самими мишенями. Это не позволяет экстраполировать результаты на другие мишени [Lapinsh и соавт., 2003].

Для построения содержательной модели необходимы сведения о достаточном количестве лигандов, активных в отношении интересующего белка. Такое требование не всегда выполнимо, особенно при исследовании недавно идентифицированной мишени.

Важно отметить, что если аффинность связывания структурно близких лигандов с одной и той же мишенью существенно различается, то такие различия могут быть обусловлены не только химической структурой, но и особенностями сайта связывания. Это зачастую может приводить к возникновению «обрывов активности» (Activity Cliff) [Guha и соавт., 2008, Wawer и соавт., 2009, Medina-Franco и соавт., 2009], когда у структурно сходных лигандов сильно различается активность в отношении одной и той мишени [Gedeck и соавт., 2006].

Сайт связывания белка-мишени обычно представляет собой подвижную структуру, позволяющую связывать различные лиганды [Surad и соавт., 2012]. В то же время связывающие карманы негомологичных мишеней, могут взаимодействовать с одними и теми же лигандами. Традиционные модели (Q)SAR не учитывают сходства между мишенями, что снижает их прогностические возможности [Cortes-Ciriano и соавт., 2015].

1.3 Построение прогностических моделей на основе структур низкомолекулярных соединений и аминокислотных последовательностей белков-мишеней. Протеохемометрическое моделирование

Чтобы преодолеть вышеуказанные ограничения был предложен подход, который Марис Лапинш и др. назвали протеохемометрическим моделированием (Proteochemometric, PCM) в 2001 г. [Lapinsh и соавт., 2001, Prusis и соавт., 2001].

Суть PCM заключается в том, что наряду с дескрипторами лиганда в модель включаются дескрипторы белков-мишеней. Таким образом, появляется возможность изучения множественных перекрестных взаимодействий белок-лиганд в обобщенном наборе данных. PCM используется для оценки белок-лигандных взаимодействий различных групп мишеней будь то наборы близкородственных белков [Lapins и соавт., 2008, Lapins и соавт., 2009, Junaid и соавт., 2010, Huang и соавт., 2012], суперсемейства [Lapinsh и соавт., 2005, Lapins и соавт., 2010], большие наборы белков, собранные без учета гомологии [Stroembergsson и соавт., 2008, Strombergsson и соавт., 2010]. Кроме того, PCM применяется для оценки взаимодействий пептид-белок [Prusis и соавт., 2013, Dimitrov и соавт., 2010, Prusis и соавт., 2008], взаимодействие антиген-антитело [Qui и соавт., 2015].

При реализации PCM для описания структур лигандов применяются те же подходы MO, что и в классическом анализе (Q)SAR. Но в этом случае белки являются не только класс-образующими признаком, а выступают также полноценными участниками взаимодействия, что требует их описания. Таким образом, вероятность связывания в PCM-моделях является функцией как от структуры лиганда, так и от структуры белка-мишени:

$$BE = \varphi(D_l, D_p)$$

где D_l – дескрипторы лиганда, D_p – дескрипторы белка, BE – оценка связывания (Binding Estimation).

Модель PCM для одного белка мишени является, по сути, моделью (Q)SAR. В этом случае результат прогноза выводится в виде бинарных величин, оценивающих возможность связывания или отсутствие таковой. При использовании PCM можно также оценить вклад отдельных дескрипторов белка в интегральную оценку его взаимодействия с лигандом.

Лапинш и соавт. показали, что в некоторых случаях PCM превосходит (Q)SAR [Lapinsh и соавт., 2001] по точности прогноза. Это было показано и другими авторами [Geppert и соавт., 2004, Ning и соавт., 2009, Paricharak и соавт., 2015]. Отметим, что в вышеуказанных исследованиях использовалась сильно упрощенная форма описания белка. Основное преимущество PCM заключается в том, что модель может описывать различные взаимодействия ряда соединений с рядом мишеней, в то же время, описывая специфические взаимодействия отдельных соединений с отдельными мишенями в наборе данных. Таким образом, PCM может эффективно соединять наборы данных (Q)SAR на основе сходства между белками-мишенями.

PCM позволяет создавать единую прогностическую модель для массива лигандов и мишеней, перекрестно соединенных между собой [Lapinsh и соавт., 2003, Lapinsh и соавт., 2005]. Подход особенно полезен в случаях, когда не удается применить трехмерное моделирование в связи с отсутствием структуры данного белка или его близкого гомолога с высоким разрешением.

1.4 Применение PCM к решению биомедицинских задач. Мишени, исследуемые при PCM-моделировании. Данные для построения моделей

Высокая точность прогноза продемонстрирована при использовании PCM-подхода в отношении нескольких групп терапевтических мишеней и их лигандов. При этом, как правило, авторы уделяли особое внимание подготовке данных, учитывая неизбежные ошибки и неточности в экспериментальных результатах,

которые могут содержаться как в исходных публикациях, так и информационных ресурсах [Kalliokoski и соавт., 2013].

Рецепторы GPCR (рецепторы, связанных с G-белками) составляют самое большое семейство белков, на которые воздействуют одобренные для клинического применения лекарства. Например, на 2017 год было известно 700 препаратов в отношении 134 белков этого семейства [Sriram и соавт., 2018]. Вероятно, по этой причине одни из первых PCM-моделей были созданы для белков этой группы и их лигандов. В ранних работах исследовались α1-адренорецепторы [Lapinsh и соавт., 2001], меланокортиновые рецепторы [Prusis и соавт., 2001, Prusis и соавт., 2002]. Позднее была выполнена работа по прогнозу связывания органических соединений с мультихимерными меланокортиновыми рецепторами [Lapinsh и соавт., 2005]. Проводились работы по прогнозу взаимодействия для GPCR, специфичных к биогенным аминам [Lapinsh и соавт., 2002b], для серотониновых, дофаминовых, гистаминовых и адренергических рецепторов [Lapinsh и соавт., 2005], а также для метаботропных рецепторов глутамата [Tresadern и соавт., 2017]. Следует отметить, что в перечисленных работах исследовались относительно небольшие наборы последовательностей, максимальное количество не превышало двадцати.

Второй по значимости группой лекарственных мишеней после GPCR являются протеинкиназы [Cohen, 2002]. Эти ферменты осуществляют пост-трансляционное фосфорилирование белков. Данная модификация задействована во многих регуляторных процессах, тогда как аномальное фосфорилирование наблюдается при различных патологиях. В настоящее время одобрено более 50 препаратов, активных в отношении протеинкиназ человека [Zhu и соавт., 2022].

Одна из первых работ по применению PCM в отношении протеинкиназ была выполнена на наборе из 317 протеинкинз, составляющих более половины кинома человека, и 38 лигандах [Lapinsh и соавт., 2010]. Эти данные представляли надежные экспериментальные результаты, поскольку были получены в рамках одного исследования, проведенного по единому протоколу одним и тем же коллективом [Karaman и соавт., 2008]. Другие авторы [Ozturk и соавт., 2018] также

использовали весьма достоверные данные, которые представляли результаты одного экспериментального исследования на 442 протеинкиназах и 68 лигандах [Davis и соавт., 2011].

Другие исследователи сосредоточились на работе с существенно большими обучающими выборками. Так выборка Кристман-Франк и соавт. [Christmann-Franck и соавт., 2016] была составлена путем досконального анализа существующих литературных и патентных данных и включала 482 протеинкиназы (более 90% кинома человека) и 2106 ингибиторов. Выборка включала достаточно разнородные данные, которые были получены на различных тест системах (Ambit/DiscoverX, Millipore, Reaction Biology и др.) и характеризовались различными показателями взаимодействия (IC50, Ki, Kd и др.). Несмотря на такую разнородность, авторам удалось построить надежные модели, обеспечивающие высокую точность прогноза. Этот же обучающий набор использовали и другие авторы [Sorgenfrei и соавт., 2018], которые дополнили его информацией из базы данных ChEMBL [Mendez и соавт., 2019]. При этом было введено дополнительное условие – не менее 20 активных и не менее 20 неактивных лигандов для каждой мишени. В результате количество протеинкиназ уменьшилось до 305, а количество лигандов увеличилось до 48953.

Были также получены модели для ядерных рецепторов, специфичных к стероидным гормонам, глюкокортикоидам, гормонам щитовидной железы. Так в работе [Qui и соавт., 2018] обучающая выборка включала 11 ядерных рецепторов и 7267 лигандов с EC50 в качестве меры аффинности.

Удалось получить надежные PCM-модели и для белковых семейств, менее изученных с точки зрения модуляции активности белков их лигандами: циклооксигеназ [Cortes-Ciriano и соавт., 2015], дигидрофолатредуктаз [Hariri и соавт., 2019], фосфодиэстераз [Rasti и соавт., 2017], ароматаз [Simeon и соавт., 2016], тимидилатсинтаз [Rasti и соавт., 2018].

Авторы вышеупомянутых подходов фокусируют свои модели на предсказании белок-лигандных взаимодействий для родственных наборов белков, в пределах таксономических групп. Другие авторы, напротив, не концентрируют

свое внимание на филогенетическом родстве белков-мишеней. Например, Ленселинк и соавт. [Lenselink и соавт., 2017] использовали внушительный массив данных, отобранный из БД ChEMBL. Он включал 1227 белков-мишеней и 204085 структур соединений. Авторы использовали весьма строгие критерии отбора, так, например, каждый белок должен быть протестирован не менее чем с 30 соединениями, по крайней мере, из двух отдельных публикаций. Значительный набор данных удалось собрать Захарову с соавт. [Zakharov и соавт., 2019]. Авторы использовали БД ChEMBL и БД Tox21 [Borrel и соавт., 2020], также применяя строгие критерии отбора на основе полей ChEMBL, отражающих надежность идентификации мишени *in vitro* и *in vivo*. Результаты, полученные при нескольких тестированиях одной и той же пары, усреднялись. Таким образом, были собраны данные для 251 998 соединений с 1082 мишенями.

Авторы используют как наиболее достоверные данные, полученные от одних и тех же исследовательских групп, так и достаточно разнородные сведения из информационных источников. В последнем случае надежность прогностической модели обеспечивается за счет фильтрации и агрегации обучающих данных. PCM-методы применялись к большому числу белковых групп, среди которых были как уже хорошо изученные лекарственные мишени (GPCR, протеинкиназы, ядерные рецепторы), так и менее изученные. Можно сделать вывод, что PCM является подходом с очень широкой областью применимости.

1.5 Подходы к описанию структур лигандов и белков-мишеней

В PCM для описания лигандов используются дескрипторы, ранее разработанные для (Q)SAR-моделирования. В то же время для включения информации о последовательности белка-мишени необходимо было реализовать специальные подходы.

Дескриптор – это численное представление молекулы или ее составных частей, которые используются в качестве входных данных при обучении и предсказании. Первые PCM-методы использовали весьма простые подходы к

описанию как белков, так и лигандов. Так, Лапинш и соавт. [Lapinsh и соавт., 2001] работали с набором производных одного соединения, что позволило представлять лиганд в виде бинарного вектора, отражающего наличие или отсутствие определенных химических групп. Белки-мишени были близкими гомологами, для характеристики которых использовали выровненные аминокислотные последовательности. Каждая из таких последовательностей была представлена вектором из ячеек, содержащих физико-химические индексы соответствующих остатков.

На практике исследователи встречаются с более сложными данными, когда белки и лиганды существенно отличаются. Мишени могут относиться к разным белковым семействам, а лиганды – к разным химическим классам. При развитии PCM исследователи предложили широкий спектр подходов, к описанию разнообразных данных о белок-лигандных взаимодействиях [Westen и соавт., 2011, Bongers и соавт., 2019].

1.5.1 Описание структур лигандов

Прежде всего, необходимо отметить, что дескрипторы лигандов могут быть построены с учетом различных структурных характеристик. Можно выделить три группы дескрипторов:

- Топологические дескрипторы представляют молекулярный граф с вершинами – атомами и ребрами – химическими связями. Это позволяет описывать атомы в контексте их окружения [Gozalbes и соавт., 2002].
- Физико-химические дескрипторы – расчетные или экспериментально определенные характеристики молекул, которые могут описывать широкий набор свойств [Raevsky и соавт., 2004].
- Трехмерные дескрипторы учитывают пространственную организацию молекулы и часто требуют дополнительной подготовки с использование подходов молекулярного моделирования [Cramer и соавт., 2007].

Полученные наборы дескрипторов различных типов затем предоставляют в виде вектора, который может содержать как бинарные, например, в случае топологических дескрипторов (рис.2), так и вещественные величины, отражающие какие-либо свойства.

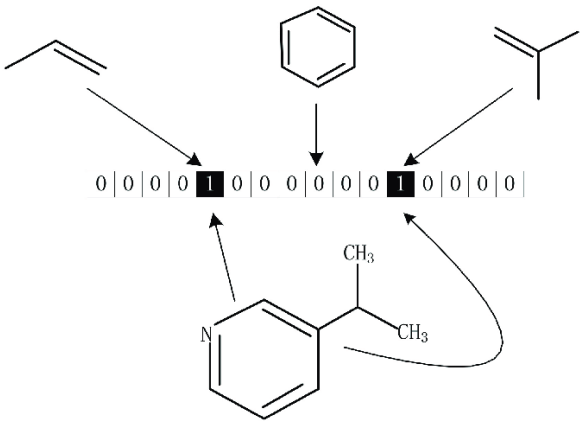

Рисунок 2. Пример генерации вектора топологических дескрипторов [Chen и соавт., 2020].

Лишь небольшое число из всего разнообразия дескрипторов, используемых в (Q)SAR для описания низкомолекулярных соединений, нашло применение в PCM при описании лигандов.

Одной из наиболее популярных систем топологических описаний, являются дескрипторы Моргана, которые учитывают ближайшие соседние атомы не далее расстояния, заданного пользователем [Morgan и соавт., 1965]. Генерация отдельного дескриптора весьма проста и включает четыре этапа:

- Каждому атому присваивается идентификатор в виде стандартного буквенного обозначения.
- Создается дескриптор атома в виде строки, отображающей как сам атом, так и соседние атомы.
- Дубликаты удаляются, при этом каждая подструктура получает целочисленный уникальный идентификатор.
- Список идентификаторов конвертируется в 2048-битный вектор (отпечаток Моргана).

Популярность отпечатков Моргана связана с быстрой генерацией дескрипторов, возможностью описать локальные особенности структуры соединений и наличием соответствующих функций в свободно доступных библиотеках [Cortés-Ciriano и соавт., 2015, Rogers и соавт., 2010, RDKit http://www.rdkit.org]. Такие дескрипторы применялись в ряде работ по PCM-моделированию [Shaikh и соавт., 2016, Simeon и соавт., 2016, Reker и соавт., 2017, Sorgenfrei и соавт., 2018, Lenselink и соавт., 2017].

В исследованиях PCM также используются физико-химические дескрипторы, такие как индекс электротопологического состояния [Mohney и соавт., 1991], параметры Липинского [Lipinski и соавт., 2001], молекулярная масса, количество ароматических связей, количество доноров/акцепторов водородных связей, коэффициент распределения октанол/вода и др. [Tresadern и соавт., 2017]. Программное обеспечение для расчета таких дескрипторов позволяет быстро получать соответствующие значения. Программные комплексы, реализованные в том числе на веб-серверах, позволяют осуществлять расчет дескрипторов для большого набора молекул. Так PaDEL обеспечивает генерацию более двух тысяч [Yap и соавт., 2011] дескрипторов, а DRAGON - почти пять тысяч [Mauri и соавт., 2006].

Некоторые авторы строят свои модели только на основе физико-химических дескрипторов [Yordanov и соавт., 2018, Qiu и соавт., 2017]. Чаще, однако, используют комбинацию топологических и физико-химических дескрипторов [Giblin и соавт., 2018, Nazarshodeh и соавт., 2018, Shar и соавт., 2016, Tresadern и соавт., 2017, Paricharak и соавт., 2015]. Следует отметить, что большое число дескрипторов не всегда дает ощутимое преимущество. Так, например, в статье Гиблина и соавт. [Giblin и соавт., 2018] показано, что отпечатки Моргана и физико-химические дескрипторы PaDEL обеспечивают близкие показатели точности при совместном применени каждого из них с дескрипторами мишени. Комбинация дескрипторов лигандов обоих типов не приводила к повышению эффективности прогноза. Обилие дескрипторов приводит к необходимости

проводить дополнительную процедуру отбора признаков [Tresadern и соавт., 2017, Nazarshodeh и соавт., 2018], что существенно усложняет разработку PCM-подходов.

Трехмерные дескрипторы используются реже, в связи с трудоемкостью их генерации. Расчет этих дескрипторов требует сложной предварительной подготовки, включающей выбор конформации лигандов [Rasti и соавт., 2018, Hariri и соавт., 2020, Giblin и соавт., 2018]. При работе с трехмерными дескрипторами необходима стандартизация химических соединений. Указанная процедура включает добавление отсутствующих или неявно заданных атомов водорода, нейтрализацию формальных зарядов, удаление солей, удаление атомов неидентифицированного типа и др. Такая оптимизация производится с помощью специализированных программ, таких как CORINA (URL:https://mn-am.com/products/corina, Amber (URL:https://ambermd.org/) и др.

К трехмерным дескрипторам относятся, в частности, GRINDS [Lapinsh и соавт., 2005, Kontijevskis и соавт., 2008, Rasti и соавт., 2017, Hariri и соавт., 2020]. Их особенность заключается в том, что они не зависят от положения молекулы в пространстве. Возможна интерпретация GRINDS-дескрипторов в отношении структурных особенностей соединения [Pastor и соавт., 2000]. На основании пространственного расположения атомов согласно выбранной конформации генерируется поле молекулярного взаимодействия, которое формируется путем автокорреляционного преобразования величин, характризующих гидрофобные контакты, доноры и акцепторы водородных связей, реакционно активные области (Рис. 3).

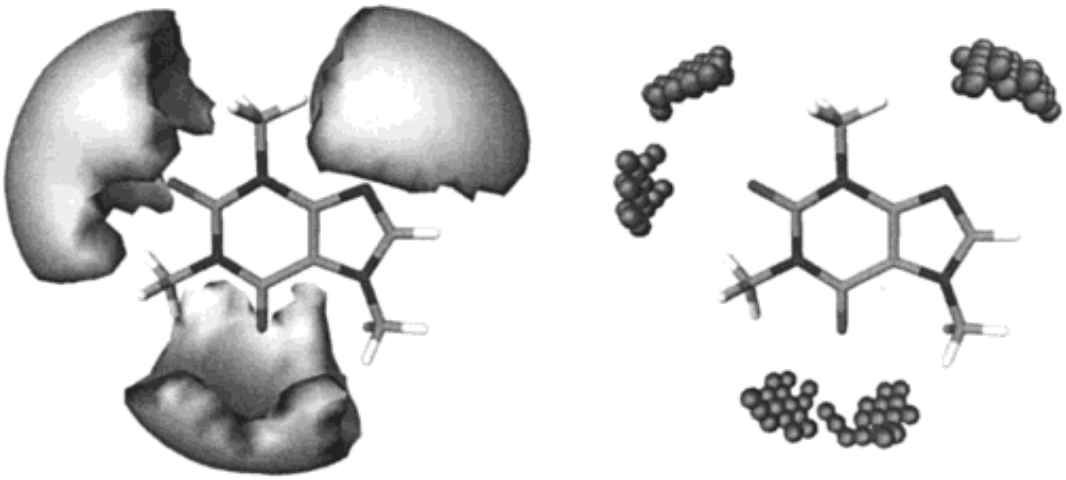

Рисунок 3. Молекула кофеина с полями молекулярного взаимодействия для амидного азота (N1) (Слева). Контур окружает точки с отрицательными значениями энергии связывания (менее -2 kcal/mol); справа молекула кофеина со 100 реакционно способными участками [Pastor и соавт., 2000].

Зачастую, авторы не приводят обоснований к выбору типа дескрипторов, а применяют, по-видимому, средства, предоставляемые доступным программным обеспечением. Тем не менее, выбор типа дескрипторов в зависимости от природы исследуемых соединений является важным шагом [Goodarzi и соавт., 2012]. Следует также отметить, что использование ограниченного числа дескрипторов дает ряд преимуществ [Murgueitio и соавт., 2012]: повышает интерпретируемость полученных моделей; снижает риск переобучения; снижает шум, связанный с избыточностью описания; позволяет получить быстродействующие и надежные модели; устраняет обрывы активности.

1.5.2 Подходы к описанию аминокислотных последовательностей белков-мишеней

При описании белка в протеохемометрике необходимо представить его также в виде совокупности дескрипторов. С этой целью применяют два основных похода. В первом случае используют предварительно выровненные аминокислотные последовательности. Во втором рассчитывают интегральные характеристики белка, такие как аминокислотный состав, аминокислотная композиция, автокорреляция и др.

При использовании множественного выравнивании, каждая последовательность представлена в виде вектора равного по размеру длине выравнивания. Для выравнивания аминокислотных последовательностей повсеместно применяются программы семейства Clustal [Dardel и соавт., 2006, Nabu и соавт., 2014, Lapinsh и соавт., 2010], несмотря на наличие более точных алгоритмов [Sviatopolk-Mirsky и соавт., 2016].

В качестве дескриптора могут использоваться как сами буквы, кодирующие аминокислотные остатки, так и различные физико-химические характеристики остатков в выровненных позициях. В случае небольшой вариабельности в определенных позициях возможно применение бинарных индексов. Так, например, исследуя пенициллинсвязывающие белки (PBP2) на длину последовательности в 581 остаток, вариабельными оказались 88 позиций [Nabu и соавт., 2014]. Авторы использовали комбинированный подход к описанию белка-мишени: 75 позиций имели только два типа остатка, для них использовался бинарный дескриптор, отражающий наличие замены отличной от эталонной последовательности (белок «дикого» типа). Если же в позициях присутствовали более двух вариантов замен их описывали с помощью физико-химических дескрипторов (Z-шкал).

Увеличение количества последовательностей нередко приводит к увеличению количества разрывов в выравнивании. Большинство позиций такого выравнивания содержат разрывы, что затрудняет их описание [Nabu и соавт., 2014]. С данной проблемой столкнулся Лапинш и соавт. [Lapinsh и соавт., 2010], которые исследовали протеинкиназы в качестве мишеней. Протеинкиазы человека представляют большую группу, включающую более 500 белков с сильно дивергировашими последовательностями [Bradley и соавт., 2021]. Проведя выравнивание, авторы удалили колонки, в которых содержалось более 50% разрывов. Итоговое выравнивание включало 263 позиции. При этом авторы не сообщают о числе удаленных позиций, тогда как подобная процедура исключения может привести к потере важной информации. Оставшиеся позиции описывались с помощью Z-шкал.

Z-шкалы являются количественными дескрипторами, полученными с помощью метода главных компонент из 26 экспериментально определенных и расчетных характеристик аминокислотных остатков. [Sandberg и соавт., 1998]. Среди экспериментальных характеристик можно отметить коэффициент удержания в семи системах тонкослойной хроматографии, а также различные значения, полученные при исследовании с помощью ядерного магнитного резонанса. Расчетных характеристик существенно больше: объем боковой цепи, logP, молекулярная масса, наличие донора/акцептора водородных связей, заряд боковой цепи, энергия высшей занятой молекулярной орбитали, энергия низшей незанятой молекулярной орбитали радикала, электроотрицательность, поляризуемость, доступность для растворителя др. Этот набор был преобразован в пять компонент, не коррелирующих друг с другом. Каждая из Z-шкал была нормализована. Возможна их интерпретация, так Z1 отражает гидрофобность, Z2 – стерические свойства, Z3 – полярность. Интерпретация Z4 и Z5 затруднительна. Наиболее доступным средство генерации таких дескрипторов является библиотека *camb* [Murrell и соавт., 2015] языка программирования *R*. Z-шкалы широко используются в PCM подходах [Lapinsh и соавт., 2010, Hariri и соавт., 2020, Tresadern и соавт., 2017, Qiu и соавт., 2017, Born и соавт., 2022].

В последнее десятилетие было расшифровано большое количество трехмерных белковых структур. Так с 2010 по 2022 год количество структур, полученных путем рентгеноструктурного анализа, возросло более чем на 100 тысяч (https://www.rcsb.org/stats/growth/growth-xray). Это не могло не повлиять на развитие PCM-методов. Используя трехмерную структуру белка-мишени, можно картировать область связывания в последовательности, а множественное выравнивание, включающее последовательности кристаллизованных белков, позволяет локализовать эти участки у всех исследуемых белков. Таким путем Гиблин и соавторы [Giblin и соавт., 2018] определяли области связывания лигандов с бромодоменами. Отбор позиций, ответственных за связывание, можно определять по их расстоянию до ближайшего атома лиганда в трехмерной структуре [Rasti и соавт., 2016, Hariri и соавт., 2018]. Некоторые авторы

используют порог в десять ангстрем [Rasti 2017], другие в пять ангстрем [Tresadern и соавт., 2017]. Для каждой группы белков-мишеней порог определяется индивидуально. Также некоторые авторы прибегают к экспертной оценке на основе подробного анализа структур [Corts-Ciriano и соавт., 2015].

Отбор анализируемых аминокислотных позиций в соответствии с трехмерной структурой позволил улучшить качество моделей для прогноза связывания протеинкиназ с ингибиторами, в сравнении моделями, построенными на основе полноразмерных последовательностей [Born и соавт., 2022]. Авторы сократили число позиций до 29 (10% от средней длины киназного домена) и использовали Z-шкалы для описания этих позиций.

Приведенные выше подходы к описанию белков показали свою эффективность в отношении семейств белков и более узких таксономических групп. При этом массовый виртуальный скрининг лигандов против белков из различных доменных семейств также является весьма актуальной задачей. Однако процедура множественного выравнивания неприменима к белкам из различных семейств, это существенно осложняет описание белков [Bongers и соавт., 2019].

Другая группа дескрипторов основана на расчете интегральных оценок аминокислотных последовательностей для наборов белков из различных доменных семейств, а также к сильно дивергировавшим семействам белков, например, к протеинкиназам [Lapinsh и соавт., 2010].

Ленселинк и соавт. [Lenselink и соавт., 2017] использовали весьма упрощенное описание белков-мишеней. Последовательности были разделены на 20 равных фрагментов (длина фрагмента зависела от длины последовательности). Для каждого фрагмента рассчитывались усредненные величины дескрипторов, которые характеризовали количество хиральных центров, заряд, акцепторы и доноры водородных связей, жесткость, ароматические связи и молекулярную массу. Затем вычислялось среднее значение для полноразмерной последовательности. Таким образом, авторы получали набор из 21×8 значений (по 20 фрагментам и один по всей последовательности). Длина последовательности была также включена в качестве отдельного дескриптора. Авторы показали, что

использование указанных дескрипторов улучшает точность прогноза по сравнению с (Q)SAR-методом, использующим только дескрипторы лигандов. Предложенный способ описания позволил привести длины векторов дескрипторов к одинаковой длине.

Многие авторы, которые работали с большими выборками, составленными из различных семейств, использовали свободно доступный веб-сервис PROFEAT [Zang и соавт., 2016] для расчета дескрипторов аминокислотных последовательностей белков-мишеней [Zakharov и соавт., 2019, Shaikh и соавт., 2016, Reker и соавт., 2017, Kim и соавт., 2020, Christmann-Franck и соавт., 2016]. Сервис позволяет рассчитать более тысячи величин, полученных на основе последовательности без использования выравнивания на основе ряда методик, описанных ниже.

Вычисление авто- и кроссковариации [Wold и соавт., 1993] позволяет описать вариацию определенных свойств (значений Z-шкал) или их комбинаций на участках последовательности различной длины. Расчет выполняется следующим образом:

$$AC_{z,lag} = \sum_{i=1}^{N-lag} \frac{V_{z,i} * V_{z,i+lag}}{N - lag}$$

$$CC_{z_a \neq z_b,lag} = \sum_{i=1}^{N-lag} \frac{V_{z_a,i} * V_{z_b,i+lag}}{N - lag}$$

где AC – автоковариация, CC – кросковариация, z – одна из пяти Z-шкал, i – аминокислотная позиция в последовательности, N – общее количество остатков в последовательности, lag – наибольшая длина Z-шкалы сопоставляемых фрагментов, V – значение Z-шкалы в позиции i. где L – максимальное значение lag.

При меньших значениях lag (например, десяти) учитывается взаимовлияние близко расположенных остатков, при больших значениях – остатков, удаленных

в первичной структуре. Количество дескрипторов, полученных при использовании пяти шкал равно 4 x L x 5^2. Используются и другие автокорреляции, например, по Морану [Horne и соавт., 1988], Гири [Geary и соавт., 1954] и нормализованная по Моро-Брото [Feng и соавт., 2000].

К дескрипторам PROFEAT также относятся и величины, рассчитанные по частотам встречаемости аминокислотных остатков и дипептидов [Chou и соавт., 2005, Gao и соавт., 2005], композиционные.

Композиционные дескрипторы PROFEAT отражают аминокислотный состав, переходы (изменение свойств по ходу) и распределение классов аминокислотных остатков [Dubchak и соавт., 1995]. Классификация остатков основана на семи свойствах аминокислотных остатков: гидрофобности, объему боковой цепи, полярности, поляризуемости, заряду, вторичной структуре и доступности для растворителя. По каждому из семи свойств определяются три класса. Так первый класс гидрофобности включает полярные аминокислотные остатки (RKEDQN), второй - нейтральные (GASTPHY) и третий гидрофобные (CLVIMFW). Дескрипторы состава представляют общий процент каждого класса в последовательности. Поскольку имеется семь атрибутов и три класса, то результирующий вектор содержит 21 значение. Дескрипторы перехода представляют частоты замен остатка одного класса на остаток другого класса по ходу последовательности - например, за остатком класса 1 следует аминокислота класса 2 или наоборот. Поскольку существует три возможных перехода между классами, длина вектора дескрипторов равна 63. Дескрипторы распределения учитывают представленность каждого атрибута в последовательности. Для каждого признака и для каждого класса вычисляются пять дескрипторов на основе следующих критериев: расположение первого остатка, первых 25 % остатков, 50 % остатков, 75 % остатков и 100 % остатков с заданным свойством. Например, если общая длина последовательности составляет N аминокислот, и все полярные аминокислоты (т.е. представители класса гидрофобности 1) входят в число первых i остатков последовательности, то дескриптор распределения для 100%

остатков данного класса будет рассчитываться как i/N. Таким образом, общее количество дескрипторов распределения составляет 5 × 7 × 3 = 121.

Дескрипторы порядка и квазипорядка PROFEAT вычисляются последовательно [Chou и соавт., 2005]. Значения порядка последовательности рассчитывается на основе физико-химических расстояний между парами аминокислот (физико-химическое расстояние по Шнайдеру-Вреде и химическое расстояние по Грантаму) [Li и соавт., 2006]:

$$\tau_d = \sum_{i=1}^{N-d} (D_{i,i+d})^2$$

где $D_{i,i+d}$ — физико-химическое расстояние между двумя аминокислотными остатками в положениях i и $i+d$ по аминокислотной последовательности, а N — длина последовательности. PROFEAT позволяет вычислять эти дескрипторы, начиная с ранга $d = 1$ (т.е. соседние остатки) до $d = 30$. После этого рассчитываются два типа т.н. дескрипторов квазипорядка на основе значений τ и аминокислотного состава белка [Li и соавт., 2006]. Дескрипторы первого и второго типа рассчитываются по следующим формулам:

$$X_a = \frac{f_a}{\sum_{a=1}^{20} f_a + w * \sum_{d=1}^{30} \tau_d}$$

$$X_d = \frac{\tau_d}{\sum_{a=1}^{20} f_a + w * \sum_{d=1}^{30} \tau_d}$$

где a – один из 20 аминокислотных остатков, f_a –частота встречаемости остатка a, w – корректирующий коэффициент $(w = 0,1)$.

Весь набор дескрипторов данного класса включает 210 значений, которые не зависят от выравнивания и содержат как количественные (физико-химические), так и качественные (буквенный код аминокислотных остатков) свойства последовательности.

Таким образом, исследователям удалось преодолеть ограничения связанные с невозможностью выравнивания негомологичных последовательностей. Однако такой способ описания приводит к существенной потере информации, которую можно получить из первичной структуры белков-мишеней.

1.5.3 Перекрестные дескрипторы для пар белок-лиганд

Для описания пар белок-лиганд предложены перекрестные (cross-term) дескрипторы [Qui и соавт., 2017]. Ряд авторов использует перемножение векторов дескрипторов лиганда и белка [Nabu и соавт., 2014, Lapinsh и соавт., 2003, Prusis и соавт., 2005, Junaid и соавт., 2010]. Такие методы успешно применяются для предсказания взаимодействующих пар, однако интерпретация вклада отдельных парных дескрипторов затруднена [Huang и соавт., 2012]. Кроме того, увеличение размеров векторов, полученных путем умножения, экспоненциально увеличивает время вычислений [Qui и соавт., 2017]. Например, в исследовании [Rasti и соавт., 2017] такой вектор имел весьма большую длину равную 25060 значений. Поэтому упомянутый подход может успешно применяться, если набор обучающих данных невелик или исходные векторы включают небольшое число дескрипторов.

Альтернативный подход основан на анализе трёхмерных структур. Каждый дескриптор представляет собой своего рода отпечаток взаимодействия, который рассчитывается из комплексов лиганд-мишень и непосредственно описывает взаимодействие лиганда с белком посредством водородных связей, ионных взаимодействий и поверхностных взаимодействий [Manoharan и соавт., 2015]. Для работы с такими дескрипторами требуется качественная трехмерная структура.

1.6 Методы машинного обучения, применяемые при PCM-моделировании

Прежде всего, необходимо отметить, что методы машинного обучения разделяются на регрессионные и классификационные. Первые позволяют прогнозировать конкретные числовые показатели взаимодействий, например, константу ингибирования, константу диссоциации и др. Классификационные же прогнозируют качественный показатель, в PCM-моделировании это чаще всего наличие или отсутствие связывания. Задавая пороговую величину показателя взаимодействия, мы делим совокупность пар белок-лиганд обучающей выборки

на класс взаимодействующих и класс невзаимодействующих. Процедура прогноза позволяет отнести исследуемую пару к одному из этих двух классов.

В качестве показателя взаимодействия используют концентрацию полумаксимального ингибирования, константу ингибирования, константу диссоциации и др. При этом пороговые значения выбирают на основе экспертной оценки или численных экспериментов. Так Захаров и соавт. использовали порог в 10 μM для различных показателей связывания [Zakharov и соавт., 2019]. Более строгий порог применяли другие авторы также для различных показателей связывания 1 μM [Shaikh и соавт., 2016] или 300 nM [Lenselink и соавт., 2017, Kim и соавт., 2020]. Для семейства протеинкиназ использовался порог в 500 nM [Sorgenfrei и соавт., 2018]. Другие авторы могут использовать интервальные значения, например, активные менее 100 nM, неактивные более 10 μM [Reker и соавт., 2017].

Среди классификационных подходов можно упомянуть алгоритм случайного леса [Breiman и соавт., 2001], при котором строится множество решающих деревьев. Решение о возможном взаимодействии принимается голосованием по большинству результатов, полученных для разных деревьев. Алгоритм случайного леса применялся во многих PCM-исследованиях [Reker и соавт., 2017, Lenselink и соавт., 2017, Shaikh и соавт., 2017, Giblin и соавт., 2018, Sorgenfrei и соавт., 2018, Merget и соавт., 2017]. Рекер и соавт. показали, что высокопроизводительные модели могут быть построены из больших наборов данных путем извлечения случайных поднаборов из обучающей выборки [Reker и соавт., 2017].

Мощным инструментом является метод опорных векторов, который также широко используется при PCM-моделировании [Shar и соавт., 2016, Giblin и соавт., 2018, Lenselink и соавт., 2017]. Основная идея заключается в переводе исходных векторов в пространство более высокой размерности, с последующим поиском разделяющей гиперплоскости.

В последние десятилетия популярность получили глубокие нейронные сети [Aliper и соавт., 2016]. Их особенность состоит в архитектуре, включающей

множество слоев, связанных большим количеством сложных связей. Внутренние слои представляют собой черный ящик, в котором информация преобразуется и взвешивается алгоритмами без какой-либо четкой обратной связи. Это затрудняет интерпретацию роли отдельных входных характеристик. Это было продемонстрировано Ленселинком и соавт., которые применили PCM-модель глубокого обучения к набору данных, полученным на основе БД ChEMBL [Lenselink и соавт., 2017].

Многие методы машинного обучения включены в свободно предоставляемые библиотеки, составленные для языков программирования (таких как R и Python) либо для платформ Weka или KNIME. Это облегчает сравнительное тестирование различных методов на предмет их эффективности. Используя библиотеки Python, Захаров и соавт. применили глубокие нейронные сети и алгоритм случайного леса для создания PCM моделей по прогнозу белок-лигандных взаимодействий [Zakharov и соавт., 2018], которые продемонстрировали достаточно высокие значения AUC – 0,84 и 0,83, соответственно. При этом работа выполнялась на больших массивах данных, представлявших в совокупности более тысячи мишеней и более 250000 соединений.

Ленселинк и соавт. получили наилучшие оценки точности при глубоком обучении (MCC = 0,57), и близкие значения для других подходов, таких как логистическая регрессия (MCC = 0,51) или случайный лес (MCC = 0,56) [Lenselink и соавт., 2017].

Из регрессионных подходов часто применяется частичная регрессия наименьших квадратов, также известная как "проекция на скрытую структуру" [Rasti и соавт., 2018, Hariri и соавт., 2020, Simeon и соавт., 2016, Lapinsh и соавт., 2010]. Этот метод объединяет метод главных компонент и множественной регрессию, что позволяет обрабатывать входные векторы с высокой степенью колинеарности.

При использовании алгоритма случайного леса для решения задач регрессии значения, предсказанные отдельными деревьями, усредняются [Cortés-Ciriano и

соавт., 2015, Christmann-Franck и соавт., 2016, Shaikh и соавт., 2016]. Во всех работах значение R^2 превышало 0,6.

При регрессии k-ближайших соседей [Nazarshodeh и соавт., 2018] усредняются наблюдения в одной и той же окрестности на пространстве признаков. Предсказание выполняется путем отнесения пары белок-лиганд к определенной окрестности. Размер окрестности задается заранее или устанавливается путем минимизации среднеквадратичной ошибки при перекрестной проверке. Для уменьшения времени расчета применяют снижение размерности с помощью лапласианской оценки [He и соавт., 2005]. Этот прием обеспечил высокую точность прогноза ($Q^2 = 0.68$, $R^2 = 0.78$) [Nazarshodeh и соавт., 2018].

Сравнительные испытания различных регрессионных подходов показали сопоставимые показатели их прогностической эффективности. Так при тестировании алгоритма случайного леса, частичной регрессия наименьших квадратов, метода опорных векторов, байесовской линейной регрессии, лучшие результаты были получены для случайного леса ($Q^2 > 0,83$) [Simeon и соавт., 2016].

Шар и соавт. сравнили по эффективности метод опорных векторов и алгоритм случайного леса. С помощью обоих методов удалось получить надежные модели, с небольшим преимуществом в пользу метода случайного леса, который обеспечил R^2, равный 0,63 против 0,61 для метода опорных векторов [Shar и соавт., 2016].

В работе [Paricharak и соавт., 2015] были проведены сравнительные испытания ряда методов, включая метод опорных векторов, градиентный бустинг и алгоритм случайного леса. При этом наибольшую оценку точности моделей получил градиенитный бустинг – $R^2 = 0,75$ и $R^2 = 0,79$ для внешней тестовой выборки. Прочие подходы показали снижение R^2 не более чем на 0,04. Сходный результат был получен при использовании Y-рандомизации [Clark и соавт., 2004] в работе [Cortes-Ciriano и соавт., 2015]. Таким образом, авторы подтвердили, что полученные результаты не зависят от случайных корреляций.

1.7 Процедура валидации PCM-моделей

Проверкой или валидацией модели называется процесс, в котором обученная модель оценивается с помощью набора тестовых данных. Тестовые данные является отдельной частью того же набора данных, из которого получен обучающий набор или же является истинно внешним набором [Wang и соавт., 2013]. С помощью валидации проверяют способность обученной модели к обобщению, оценивают различные подходы к сбору данных, сравнивают способы описания исследуемых объектов [Alpaydin 2010]. Кроме того, процедура валидации направлена на поиск алгоритма с наилучшей производительностью.

При проведении PCM-исследований крайне желательно наличие внешней обучающей выборки. В случае ее отсутствия применяют различные варианты перекрестной валидации. Суть этой процедуры состоит в том, что часть обучающей выборки исключается и обучение проводится без нее. Обученная модель проходит тестирование на оставшихся данных. Такой процесс может повторяться итеративно. Ряд авторов [Zakharov и соавт., 2018, Гиблин и соавт., 2018, Shar и соавт., 2016, Paricharak и соавт., 2015] проводили пятикратное разделение обучающей выборки. Каждый раз обучение проводилось на 80% выборки, а тестирование на оставшихся 20%. В работе [Cortes-Ciriano и соавт., 2015] использована условно внешнюю тестовую выборку. Исходный набор был разделен на шесть частей, одна из которых служила внешней тестовой выборкой. Остальные пять частей использовались для перекрестной валидации с целью оптимизации параметров модели.

Ленселинк и соавт. провели случайное разбиение выборки на 70% обучающих данных и 30% тестовых [Lenselink и соавт., 2017]. В той же работе авторы также применили разбиение данных с учетом времени их поступления в информационный ресурс. Так что часть данных, которая была внесена ранее, составила обучающую выборку, а внесенные позднее – тестовую выборку.

Другой принцип составления внешней выборки основан на кластеризации обучающей выборки по методу K-средних [Simeaon и соавт., 2016]. Из каждого кластера часть пар белок-лиганд были отобраны для тестовой выборки, которая

составила 20% от исходного набора. Авторы не сообщили о том, исключалась ли одновременно и мишень и лиганд из обучающей выборки.

Вышеуказанный недостаток отмечен нами в другой работе [Hariri и соавт., 2020]. PCM-моделирование проводилось на наборе из двух белков и 284 низкомолекулярных соединений, авторы сообщают, что в качестве внешней выборки из данных были исключены 20% данных. Из текста статьи нельзя понять проводилось ли исключение белков из обучающей выборки. В другой работе [Rasti и соавт., 2018] исключению при пятикратной валидации повергались только лиганды, тогда как мишени оставались как в обучающей, так и в тестовой выборке. Такую же особенность можно обнаружить и в других статьях [Hariri и соавт., 2018, Rasti и соавт., 2017].

Вышеуказанный способ перекрестной валидации без явного исключения мишеней обусловил, по всей вероятности, отсутствие принципиальных различий в точности прогноза при включении дескрипторов белков и лигандов по сравнению с (Q)SAR-методами, учитывающими дескрипторы лигандов [Lapinsh и соавт., 2003, Freyhult и соавт., 2005]. Подобные работы в связи с особенностями их тестирования не могут в полной мере служить для оценки эффективности PCM. Наличие одних и тех же объектов в обучающей и тестовой выборках может рассматриваться как разновидность самоузнавания и искусственно завышать оценку точности. При этом, когда авторы в явном виде отражают факт полного исключения как лиганда, так и белка отмечается снижение точности [Sorfengei и соавт., 2018, Shi и соавт., 2018].

Важно подчеркнуть, что большинство авторов четко не отражают стратегию валидации. Более того, один из родоначальников метода, также не формулирует, как валидировать такие подходы [Lapinsh и соавт., 2001]. Отмечая, при этом, что ключевой особенностью является возможность предсказания лигандов для новых мишеней. Методика валидации (Q)SAR-методов отличается от таковой при PCM, так как (Q)SAR не предполагает прогноз для несуществующего класса (то есть мишени, которой нет в обучающих данных). Ряд авторов вносит ясность в этот

вопрос [Cortes-Ciriano и соавт., 2015, Sorgenfrei и соавт., 2018, Paricharak и соавт., 2015], разделяют подходы к валидации на сценарии прогноза (рис. 4).

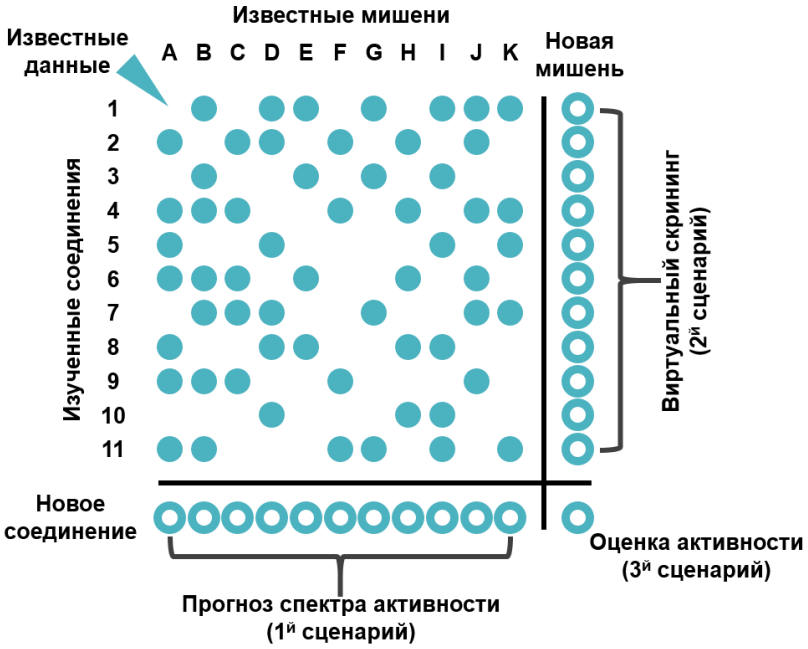

Рисунок 4. Сценарии, реализуемые при PCM-моделировании.

Согласно первому сценарию, построение модели осуществляется только на основе структуры низкомолекулярного соединения, а тестирование осуществляется для структуры нового лиганда, среди белков-мишеней с установленным спектром лигандов. Этот сценарий соответствует подходу (Q)SAR. Согласно второму сценарию, обучение осуществляется только на основе аминокислотной последовательности, прогноз осуществляется для новой последовательности среди лигандов с уже установленным спектром мишеней. Центральной же задачей PCM является прогнозирование взаимодействий для ситуации, когда ни лиганд, ни белок-мишень не аннотированы с точки зрения их взаимодействий. Это отражено в третьем сценарии, когда на вход подается новый лиганд и новый белок-мишень. При этом при третьем сценарии отмечается ожидаемое снижение точности в связи с тем, что при валидации с исключением обоих компонентов происходит обеднение обучающих данных [Cortes-Ciriano и соавт., 2015, Sorgenfrei и соавт., 2018]. В то же время точность прогноза и в этом случае остается достаточно высокой.

1.8 Заключение

Анализ актуальных публикаций по теме PCM-моделирования показал, что данный подход обладает широкой областью применимости. Авторам удалось разработать надежные методы, успешно примененные для разнообразных биологических мишеней и их лигандов. Обучающие данные включали от сотен до сотен тысяч взаимодействующих пар. Даже для небольших относительно бедных обучающих наборов были получены надёжные результаты.

Для описания структур низкомолекулярных лигандов используются дескрипторы, широко применяемые для (Q)SAR-исследований, которые продемонстрировали свою эффективность и при решении задач PCM-моделирования. Обширный инструментарий для генерации основных типов дескрипторов для химических соединений находится в свободном доступе и позволяет быстро проводить генерацию наиболее популярных молекулярных дескрипторов. Методы, основанные на трехмерной структуре лиганда, требуют ручного вмешательства на основных этапах подготовки и поэтому находят ограниченное применение.

До развития PCM-моделирования такого массового применения последовательностей для задач прогнозирования с помощью машинного обучения не отмечалось. Это повлекло за собой развитие подходов к описанию белков, основные направления которых рассмотрены в разделе «Описание белковых последовательностей». Как следует из анализа литературы, подготовка последовательностей весьма разнородна по своим методикам. Выбор системы дескрипторов зависит от количества последовательностей, их дивергенции, а также доменного состава. Описание белков может быть представлено бинарными векторами, интегральными характеристиками на основе физико-химических свойств белка или его фрагментов, а также с помощью достаточно трудоемкого анализа трехмерных структур. Выбор способа описания белка существенно влияет на эффективность прогноза. Часто при смене группы белков-мишеней исследователи вынуждены модифицировать свои подходы, и это прежде всего связано с особенностями описания белков. Таким образом, подготовка

аминокислотных последовательностей часто оказывается решающим фактором при создании PCM-моделей. Назрела необходимость в создании нового более универсального метода с широкой областью применимости, который бы позволил анализировать взаимодействия лигандов с белками различных таксономических групп.

В PCM применяются практически все наиболее распространенные методы машинного обучения. Сравнительные тесты показали, что значения точности прогноза, полученные с помощью разных методов, различаются незначительно. Поэтому при выборе метода на первый план выходит быстрота разработки и скорость расчетов.

Особое внимание при оценке эффективности PCM-методов следует уделить корректной валидации, которая должна проводиться с учетом возможных сценариев, которые реализуются при компьютерной оценке белок-лигандных взаимодействий.

ГЛАВА 2. МАТЕРИАЛЫ И МЕТОДЫ

2.1 Информационные источники

В данной работе мы использовали разнородные информационные ресурсы для более полного описания белок-лигандных взаимодействий. Собранные данные отражали различные аспекты белок-лигандных взаимодействий и были использованы для обучения PCM-моделей и оценки точности прогноза. Ниже представлено описание основных источников данных, использованных в исследовании.

- Информация о взаимодействии белков с низкомолекулярными лигандами получена из БД ChEMBL (https://www.ebi.ac.uk/chembl/) [Mendez и соавт., 2019]. Это один из крупнейших свободно доступных ресурсов, посвященных биологической активности химических соединений. В нем содержится информация о более чем двух миллионах химических соединений. Сведения получены из разнообразных источников, включая результаты массового скрининга и отдельных исследований, патенты и др. Указаны данные об использованных методах и экспериментальных условиях: взаимодействие белка и лиганда *in vitro*, испытания на клеточных линиях, тесты на лабораторных животных и др. Каждая запись содержит структурную формулу соединения, количественную или качественную оценку результата в зависимости от измеряемого показателя. Обновление БД осуществляется не реже раза в год. В работу включены результаты, полученные на основе версии 28 (февраль 2021 года). Прежде чем использовать данные ChEMBL в работе, необходимо было произвести их предподготовку: отфильтровать по степени надежности, и свести оценки взаимодействия белок-лигнад, полученные для различных показателей, к единой бинарной форме.
- Стэндфордская база данных содержит информацию о лекарственной устойчивости вариантов вируса иммунодефицита человека (ВИЧ) [Shafer и соавт., 2006] (https://hivdb.stanford.edu/). Записи содержат сведения о лекарственной чувствительности мутантных вариантов трех белков ВИЧ: обратной

транскриптазы, протеазы и интегразы. Большинство последовательностей имеет небольшое число замен относительно «дикого» типа. Поэтому каждая из трех групп представлена в виде множественного выравнивания с указанием замен. Последовательности, содержащие смеси остатков в одной и той же позиции исключались в связи с невозможностью точного восстановления исходных вариантов. Количественная оценка эффектов, обусловленных мутациями, выражается в виде отношения резистентности (Resistance Ratio). Эта величина отражает то, насколько у штамма, экспрессирующего соответствующий вариант белка, снижается чувствительность к рассматриваемому препарату. Информация о резистентности штаммов ВИЧ содержится для 22 препаратов: Абакавир, Атазанавир, Биктегравир, Дарунавир, Диданозин, Долутегравир, Зидовудин, Индинавир, Ламивудин, Лопинавир, Невирапин, Нелфинавир, Ралтегравир, Рилпивирин, Саквинавир, Ставудин, Тенофовир, Типранавир, Фосампренавир, Элвитегравир, Этравирин, Эфавиренз.

- База знаний UniProtKB представляет собой крупнейший информационный ресурс, посвященный белкам [Pundir и соавт., 2017]. Записи содержат аминокислотные последовательности и их подробные функциональные аннотации. Использовались данные из SwissProt – раздела, который содержит наиболее достоверные записи, курируемые экспертами. UniProtKB интегрирован с другими информационными ресурсами по белкам, например, с Pfam и PDB, к которым мы обращались в данной работе.
- БД Pfam [Mistry и соавт., 2021] содержит данные о семействах, каждое их которых объединяет группу гомологичных доменов со сходной пространственной структурой и сходными функциями. Большинство белков человека содержат, как минимум, последовательность одного домена, описанного в Pfam. При отборе последовательностей, отнесенных к доменным семействам, использовались перекрестные ссылки UniProtKB и Pfam.
- Protein Data Bank (PDB) основной свободно доступный ресурс аккумулирующий все трехмерные структуры белков, разрешенные с помощью

рентгеновской кристаллографии, ЯМР-спектроскопии и других методов [Velankar и соавт., 2021]. Существенная часть структур расшифрована в комплексе с лигандами. Мы использовали такие записи для проверки результатов, полученных при предсказании функционально важных позиций. При картировании лиганд-связывающих аминокислотных остатков мы обращались к базе данных PDBbind, которая приводит сведения по трехмерным структурам для пар белок-лиганд с указанием показателей аффинности. Последние почерпнуты из разных источников, включая ChEMBL (http://www.pdbbind.org.cn/) [Zhihai и соавт., 2015].

- Информация о белковых доменах, представленных в разрешенных трехмерных структурах, извлекалась на веб-сервисе SIFTS [Dana и соавт., 2019]. Это позволило отбирать структуры с интересующими нас доменами.
- Поскольку лигандам в записях PDB присвоены трехбуквенные обозначения, для установления идентификаторов, используемых в ChEMBL (InChI), мы использовали ресурс Ligand Expo (http://ligand-expo.rcsb.org/) [Rose и соавт., 2013].

2.2 Программные средства и языки программирования

Скрипты для обработки данных о взаимодействии белков и лигандов, стандартизации структур лигандов и форматировании данных были написаны с использованием интерпретируемого языка программирования Python [Yogesh и соавт., 2019] в среде разработки PyCharm. Эти средства были выбраны в связи с тем, что стандартная библиотека Python включает большой набор разнообразных функций по обработке текста и таблиц.

Для обработки и анализа данных была использована платформа KNIME, которая реализует принцип «визуального» программирования с помощью графических модулей (рис. 5) [Mazanetz и соавт., 2012]. Так модуль «Read csv» осуществляет чтение файла соответствующего формата, а модуль «Sort» осуществляет сортировку, указанных пользователем столбцов. Использование стандартных модулей KNIME несколько ограничивает возможности по сравнению с «обычным» программированием. Однако использование этой

платформы существенно облегчает работу с таблицами за счет снижения риска, связанного с погрешностями при написании кода вручную.

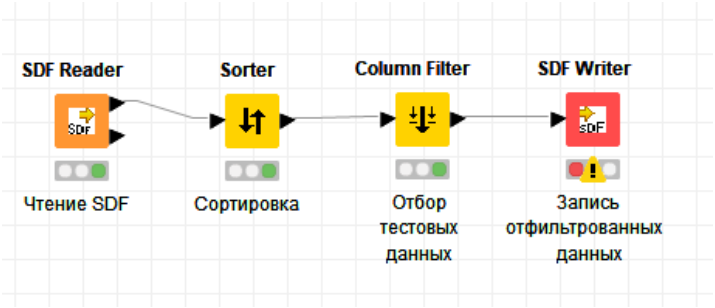

Рисунок 5. Пример визуального программирования на платформе KNIME.

Для картирования результатов работы программы SPrOS в позиционном режиме использовалась программа визуализации молекул PyMOL. Данная программа позволяет отображать структуры, полученные из PDB. Встроенные функции позволяют писать скрипты на языке программирования Python, например, для удобного и быстрого картирования предсказанных аминокислотных остатков в структуре белок-лигандного комплекса. Следует также упомянуть функцию трехмерного выравнивания PyMOL, которая была использована для сопоставления конформаций одного и того же белка.

2.3 Метод PASS. Обработка структур лигандов

Прогноз белков-мишеней для лигандов (SAR-прогноз) был сделан с использованием метода PASS (Prediction of Activity Spectra for Substances), реализованный в одноименной программе [Filimonov и соавт., 2014]. Программа PASS разработана в Лаборатории структурно-функционального конструирования лекарств ИБМХ им. В.Н. Ореховича. Как и в других (Q)SAR-методах, главный принцип, положенный в основу PASS состоит в том, что свойства низкомолекулярного соединения рассматриваются как функция от его структуры. PASS осуществляет качественный прогноз, то есть обучающие данные должны иметь категориальные значения, в нашем случае есть ли связывание белка-мишени с лигандом или нет.

Для описания структур низкомолекулярных лигандов используются дескрипторы многоуровневых атомных окрестностей – MNA (Multilevel

Neighborhoods of Atoms) [Filimonov и соавт., 1999]. С помощью MNA-дескрипторов осуществляется описание лигандов на основе взаимного расположения атомов без учета типа связи между ними. Положение атомов водорода восстанавливается согласно валентностям и зарядам других атомов. Дескриптор представляет собой атом-центрированную запись фрагментов лиганда. Дескрипторы строятся рекурсивно для каждого атома, так, что дескриптор MNA 0-го уровня является меткой самого атома «*A*»; дескриптор MNA следующего уровня, записывается как – $A(D_1D_2...D_i...)$, где D_i – дескриптор MNA предыдущего уровня для *i*-го непосредственного соседа данного атома с меткой A. В используемой версии PASS использовались два уровня дескриптора – MNA1 и MNA2 (рис. 6).

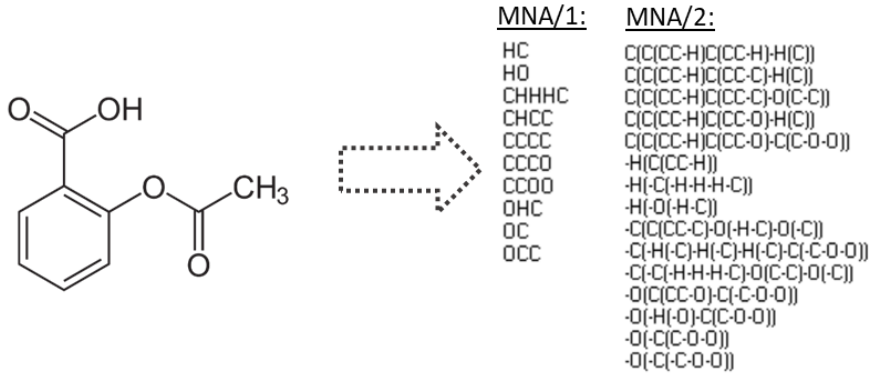

Рисунок 6. Список MNA-дескрипторов для молекулы аспирина.

Структура лиганда должна удовлетворять определённым условиям:

- атомы должны быть записаны с помощью символов периодической системы Менделеева;
- связи могут быть только ковалентные (ординарные, двойные, тройные);
- структура должна содержать как минимум три углеродных атома;
- молекула должна быть электронейтральной;
- структура должна содержать только одну молекулу;
- изолированные атомы недопустимы;
- молекулярная масса не более 1250 а.е.;
- структура не может содержать более четырех пептидных связей.

Такой способ описания не учитывает стереоизомерии лигандов, однако в большинстве случаев подобная информация недоступна. Стереоизомеры одного и того же соединения характеризуются одинаковым набором MNA-дескрипторов.

Предсказание в PASS основано на модифицированном наивном Байесовском классификаторе [Filimonov и соавт., 2008], который сопоставляеят локальные фрагменты структур лигандов, представленных в виде MNA-дескрипторов. В результате прогноза рассчитываются вероятностные оценки связывания лиганда с той или иной мишенью обучающей выборки. Две величины на выходе P_a и P_i отражают, соответственно, вероятность связывания с белком-мишенью и вероятность, того что связывания не произойдет. Если P_a превышает P_i, то прогноз считается положительным, и отрицательным в противном случае.

2.4 Метод SPrOS, реализация PCM-моделирования

Для обработки последовательностей, прогноза на основе аминокислотной последовательности, а также PCM-моделирования использовалась программа SPrOS (Specificity Projection On Sequence). Подход разработан в Лаборатории структурно-функционального конструирования лекарств ИБМХ им. В.Н. Ореховича [Alexandrov и соавт., 2008, Karasev и соавт., 2020].

Описание аминокислотных последовательностей осуществляется на основе оценок, отражающих сходство аминокислотных позиций тестовой и обучающей последовательности в контексте локального окружения (рис. 7). Так, аминокислотная последовательность тестового белка пофрагментно сравнивается с последовательностями обучающей выборки.

Пусть тестовая последовательность белка (*T*) состоит из перекрывающихся фрагментов *[1; F], [2; F+1] …[l - F+1; l]*, где *l* – длина последовательности, а *F* – длина сопоставляемых фрагментов. В результате смещения тестовой последовательности относительно последовательности обучающей выборки *K* каждая пара фрагментов, совмещенных при данном сдвиге, получает оценку *R* как сумму оценок сходства совмещенных позиций из последовательностей *T* и *K*:

$$R_{ih} = \sum_{j=i}^{j+F-1} sim\left(T_{j,} K_{j+h}\right);$$

где i – первая позиция фрагмента последовательности T, h – величина сдвига между последовательностями T и K, F - длина сопоставляемых фрагментов (на рис. 7 для наглядности $F = 20$), $sim(T_a, K_b)$ – мера сходства совмещенных остатков - в данной работе идентичность. При оценке точности прогноза по второму и третьему сценариям прогностического режима значение параметра ***F*** устанавливалось по умолчанию равным 30.

```
AANRDPSQFPDPHRFDVTRDTRGHLSFGQGIHFCMGRPLAKLEGEVA  2
ANRDPSQFPDPHRFDVTRDTRGHLSFGQGIHFCMGRPLAKLEGEVAL  1
NRDPSQFPDPHRFDVTRDTRGHLSFGQGIHFCMGRPLAKLEGEVALR  1
RDPSQFPDPHRFDVTRDTRGHLSFGQGIHFCMGRPLAKLEGEVALRA  0
DPSQFPDPHRFDVTRDTRGHLSFGQGIHFCMGRPLAKLEGEVALRAL  1
PSQFPDPHRFDVTRDTRGHLSFGQGIHFCMGRPLAKLEGEVALRALF  2
SQFPDPHRFDVTRDTRGHLSFGQGIHFCMGRPLAKLEGEVALRALFG  1
QFPDPHRFDVTRDTRGHLSFGQGIHFCMGRPLAKLEGEVALRALFGR  1
FPDPHRFDVTRDTRGHLSFGQGIHFCMGRPLAKLEGEVALRALFGRF  2
PDPHRFDVTRDTRGHLSFGQGIHFCMGRPLAKLEGEVALRALFGRFP  0
DPHRFDVTRDTRGHLSFGQGIHFCMGRPLAKLEGEVALRALFGRFPA  1
PHRFDVTRDTRGHLSFGQGIHFCMGRPLAKLEGEVALRALFGRFPAL  0
HRFDVTRDTRGHLSFGQGIHFCMGRPLAKLEGEVALRALFGRFPALS  9
RFDVTRDTRGHLSFGQGIHFCMGRPLAKLEGEVALRALFGRFPALSL  0
FDVTRDTRGHLSFGQGIHFCMGRPLAKLEGEVALRALFGRFPALSLG  3
DVTRDTRGHLSFGQGIHFCMGRPLAKLEGEVALRALFGRFPALSLGI  1
VTRDTRGHLSFGQGIHFCMGRPLAKLEGEVALRALFGRFPALSLGID  1


GTAINKPLSEKMMLFGMGKRRCIGEVLAKWEIFLFLAILLQQLEFSV  9
```

Рисунок 7. Сдвиг тестовой последовательности относительно обучающей позволяет найти пару наиболее похожих фрагментов. Нижняя строка – тестовая последовательность. Верхние строки представляют собой последовательные сдвиги одной и той же обучающей последовательности. Остатки, совпавшие при разных сдвигах, выделены. Количество совпавших остатков указаны справа. Наиболее сходный фрагмент выделен рамкой.

Каждой позиции присваивается наибольшая из оценок R_{ih}, полученных при всех допустимых смещениях для всех фрагментов последовательности T, включающих эту позицию:

$$S_p = \max_{h,i} R_{ih}, p - F < i \le p$$

Таким образом, тестовая последовательность сравнивается со всеми *k*-последовательностями обучающей выборки, в результате чего каждая позиция тестовой последовательности получает набор позиционных оценок S_{pk}.

При прогнозе взаимодействия тестового белка с *C*-лигандом каждый *k*-белок обучающей выборки относится к классу мишеней в соответствии с заданным коэффициентом $a_k(C)$. Принадлежность того же *k*-белка к невзаимодействующим белкам определяется коэффициентом $b_k(C)$. Мы использовали бинарные значения $a_k(C)$ и $b_k(C)$, когда прогноз выполнялся по второму сценарию без учета структурных характеристик лигандов (рис. 4). При PCM-моделировании (третьем сценарии) коэффициентам $a_k(C)$ и $b_k(C)$ присваивались вещественные значения, рассчитанные на основе структурного сходства лигандов с помощью метода PASS.

Позиционные оценки вместе с коэффициентами принадлежности составляют основу расчета оценок взаимодействия белка при двух режимах прогноза, предусмотренных в программе SPrOS – прогностическом и позиционном.

Прогностический режим позволяет получить интегральную оценку взаимодействия белок-лиганд с помощью следующих вычислений. Для каждой *p*-позиции тестовой последовательности рассчитывается величина t_p с учетом всех оценок, полученных при сопоставлении со всеми *N* белками обучающей выборки:

$$t_p = \frac{\sum_{k=1}^{N} S_{pk} \times \left[a_k(C) - b_k(C)\right]}{\sum_{k=1}^{N} S_{pk} \times \left[a_k(C) + b_k(C)\right]},$$

Позиционные оценки усредняются по всей длине последовательности:

$$t = \sin\left[\frac{1}{m}\sum_{p=1}^{m} \arcsin(t_p)\right]$$

где *m* – длина последовательности.

Для учета представительности класса в обучающей выборке вычисляется величина t_0:

$$t_0 = \frac{\sum_{k=1}^{N}[a_k(C) - b_k(C)]}{\sum_{k=1}^{N}[a_k(C) + b_k(C)]}$$

Интегральная оценка *BE* для пары белок-лиганд может варьировать от «-1» (наиболее вероятно отсутствие связывания) до «1» (наиболее вероятно связывание) и определяется по формуле:

$$BE(C) = \frac{t - t_0}{1 - tt_0}$$

Позиционный режим, позволяющий оценить вклад отдельных аминокислотных позиций в специфичность связывания, описан в разделе 2.6.

2.5 Сценарии прогностического режима

Основным преимуществом PCM-моделирования является возможность прогнозировать связывание для пар белок-лиганд, у которых не аннотирован спектр взаимодействий ни у одного компонента (рис.4, рис. 8). Также существует возможность осуществлять прогноз только на основе структуры лиганда, такой вариант прогнозирования мы будем называть первым сценарием, он соответствует классическому (Q)SAR-моделированию. При этом предсказание осуществляется только для белков-мишеней с установленным лигандным спектром. В таком сценарии белки-мишени выполняют роль класс-образующих признаков. С другой стороны, возможно предсказание только на основе аминокислотных последовательностей, в таком случае уже лиганды используются как класс-образующие признаки. При этом прогноз осуществляется только для лигандов с уже известным спектром белков-мишеней.

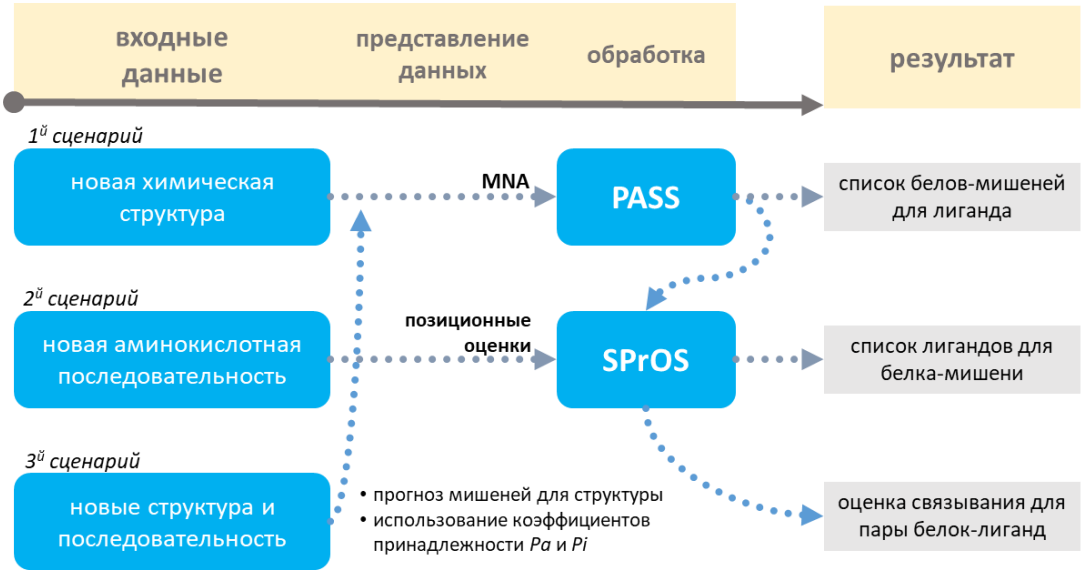

Рисунок 8. Схема работы разработанного подхода.

Технически первый сценарий реализован с помощью метода PASS. Второй сценарий реализован с помощью программы SPrOS. При реализации третьего сценария (PCM), происходит объединение процедур первого и второго сценариев (рис. 8). Таким образом, происходит оценка соответствия между белком-мишенью и лигандом, причем каждый оба метода основаны на выявлении локальных особенностей каждого из компонентов, обуславливающих белок-лигандные взаимодействия.

При валидации метода на первом этапе с помощью программы каждый лиганд *C* обучающей выборки сопоставляется со всеми лигандами *k*-того белка за исключением самого лиганда *C*. Предсказанные оценки взаимодействия P_a и P_i служат в дальнейшем как коэффициенты $a_k(C)$ *и* $b_k(C)$, характеризующие возможное взаимодействие *k*-того белка и лиганда *C*.

2.6 Позиционный режим работы программы SPrOS

При исследовании белок-лигандных взаимодействий полезно иметь информацию о тех позициях в последовательности белка, на основе которых был осуществлен прогноз. Такие задачи решаются в позиционном режиме программы SPrOS. Как и в прогностическом режиме на первом этапе работы алгоритма осуществляется пофрагментное сравнение тестовой последовательности со всеми

последовательностями обучающей выборки (рис. 7). При оценке точности предсказания в позиционном режиме в значение параметра *F* устанавливалось равным 7. Позиционные оценки усредняются с учетом величин $a_k(C)$ и $b_k(C)$, которые представляют собой соответственно коэффициенты принадлежности белка к классу мишеней лиганда *C* и дополнению этого класса, объединяющему невзаимодействующие белки:

$$a_p = \frac{\sum_k S_{pk} a_k(C)}{\sum_k a_k(C)} \qquad b_p = \frac{\sum_k S_{pk} b_k(C)}{\sum_k b_k(C)}$$

Оценка специфичности позиции *p* к лиганду *C* рассчитывается следующим образом:

$$E_{pa} = \frac{a_p - b_p}{a_p + b_p}$$

Оценка E_{pa} варьирует от -1 (нет специфичности к лиганду) до +1 (специфичность к лиганду).

Для оценки статистической значимости результатов при позиционном режиме прогноз многократно повторяется для последовательностей, группируемых случайных образом при сохранении размеров классов. На основе полученного распределения оценок определяются величины *p*-уровня значимости для оценок E_{pa}, рассчитанных при исходном распределении.

Для оценки вклада аминокислотных остатков в формировании белок-лигандного комплекса разработан отдельный веб-сервис (http://www.way2drug.com/spros/), использующий метод SPrOS в позиционном режиме.

2.7 Оценка эффективности разработанного подхода

При оценке эффективности прогноза, основанном на описании обоих компонентов взаимодействия, необходимо определить процедуру валидации. Если вероятность связывания предсказывается только по структурам лигандов

или только по последовательностям белка (первый или второй сценарий), то значимость результата рассчитывается путем скользящего контроля с исключением по одному (Leave-One-Out Cross Validation, LOO CV). В этом случае на каждом шаге итеративной процедуры один из объектов исключается из выборки и используется как тестовый, а обучение осуществляется на оставшейся части выборки. При PCM-моделировании необходимо, чтобы элементы каждого из двух типов последовательно исключались и использовались как тестовые. Только в таком случае можно объективно оценить прогностическую эффективность подхода. В настоящей работе валидация данных по лигандам проводилась параллельно с расчетом коэффициентов взаимодействия. Эти же величины подставлялись в качестве коэффициентов при валидации с учетом последовательностей мишеней по третьему сценарию.

Мы определяли эффективность с помощью критерия инвариантной точности прогноза (Invariant Accuracy of Prediction) [Filimonov, Poroikov 2008]. Каждая из оценок Sc_a, полученных для каждого из N_a объектов предсказываемого класса сопоставляется с каждой из оценок Sc_b, полученных для N_b белков дополнения, то есть не входящих в указанный класс. Общее число пар $< Sc_a, Sc_b >$ равно произведению размеров обеих групп $N_a \times N_b$. Если подсчитать число правильно классифицированных пар, то есть таких, в которых оценка для класса больше оценки для некласса $N(Sc_a > Sc_b)$, то следующая формула

$$IAP = \frac{N(Sc_a > Sc_b)}{N_a N_b}$$

определяет вероятность того, что любая пара $<Sc_a, Sc_b>$, взятая наугад из результатов будет «правильной».

Величина IAP численно равна площади под ROC-кривой, которая часто используется как показатель точности прогноза.

Позиционный режим работы алгоритма валидируется иначе. На сегодняшний день трудно представить белок, у которого на основе эксперимента были бы проаннотированы все аминокислотные остатки, специфичные в

отношении какой-либо функции. Поэтому в данном случае мы использовали два варианта тестирования [Karasev и соавт., 2016]. В первом случае валидация проводилась на группе искусственно сгенерированных последовательностей с заменами, специфичными для отдельных подгрупп. Во втором случае тестирование поводилась на последовательностях реальных белков с разметкой остатков, контактирующих с лигандами в трехмерных структурах. Названия лигандов же при этом используются как класс-образующие признаки.

2.8 Генерация искусственных последовательностей

Искусственные аминокислотные последовательности были созданы путем моделирования биологической эволюции. В качестве исходной последовательности был использован N-концевой участок (225 остатков) цитохрома P450 2C9 (UniProt AC P11712). Замены контролировались на основе стандартного генетического кода и частот использования кодонов.

После заданного уровня итераций последний набор последовательностей рассматривался как модельное белковое семейство, для которого строилось филогенетическое дерево. Далее набор последовательностей случайным образом разбивался на группы, которые не совпадали с филогенетическим делением. В выбранные участки аминокислотных последовательностей вводились специфичные для каждой группы замены.

ГЛАВА 3. РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

3.1 Подготовка тестовых и обучающих данных. Область применимости обучающих данных

Прогноз белок-лигандных взаимодействий осуществляется на массивах данных, различающихся по размеру и гетерогенности.

Прогноз с использованием крупных выборок, которые содержат сведения по белкам из разных семейств и лигандам из разных химических, весьма полезен на начальных этапах исследования. При этом разные типы данных представлены весьма неоднородно. Хорошо известные лекарственные мишени и их лиганды часто представлены в гораздо большей степени, чем новые объекты исследования.

При работе с гомологичными белками выборка, как правило, более однородна. В этом случае следует ориентироваться на доменные семейства, поскольку мишени одной и той же группы часто являются компонентами многодоменных белков, весьма отличающихся по составу. Так тирозиновые протеинкиназы могут быть отдельными белками или входить в состав рецепторов, проявляя при этом сродство к идентичным или структурно-близким ингибиторам. Задача группировки мишеней облегчается тем, что современные классификации белков основаны на сходстве доменных последовательностей [Paysan-Lafosse и соавт., 2023]. Кроме того, при составлении выборок с мишенями из одного семейства легче унифицировать экспериментальные данные по взаимодействию с лигандами, поскольку соответствующие результаты получены с помощью сходных протоколов тестирования. Прогноз белок-лигандных взаимодействий на выборках, представляющих родственные белки и их лиганды, производится на более поздних этапах виртуального скрининга.

При работе с близко гомологичными мишенями можно ожидать наиболее точные результаты. В качестве характерного примера можно привести мутантные формы вирусного белка. При этом доступны представительные результаты тестирования с одним и тем же набором лигандов, зачастую проведенные в одних и тех же условиях.

Исходя из этих соображений, мы собрали три типа данных для отладки и валидации разработанного нами подхода.

В первом случае, набор данных представлял сведения по белок-лигандным взаимодействиям, собранные из БД ChEMBL без учета классификации белков по семействам.

Во втором случае, исследовались четыре набора данных из того же источника, сгруппированные по структурно-функциональным группам белков. Три группы протеинкиназы, GPCR (рецепторы, связанные с G-белком) и ядерные рецепторы представляли три разных семейства белков-гомологов. Группа ионных каналов включала белки из нескольких белковых семейств. Выбор этих групп семейств был обусловлен представительностью данных и тем, что включенные в них белки являются перспективными лекарственными мишенями [Santos и соавт., 2017].

Для исследования эффективности подхода на мутантных формах белков (близких гомологов) мы использовали «Stanford HIV database», которая содержит сведения по тестированию антивирусных препаратов на взаимодействие с мутантными вариантами трех белков ВИЧ [Rhee и соавт., 2003].

Указанные три группы данных являются основными областями применимости PCM-методов.

Для тестирования позиционного режима на основе экспериментальных данных мы использовали наиболее надежную выборку, которая представляла результаты одного исследования [Karaman и соавт., 2008].

3.2 Подготовка обучающих данных из БД ChEMBL

Сбор данных из БД ChEMBL проводился с использованием процедуры фильтрации, что позволило отобрать наиболее надежные данные. Достоверность оценивалась по содержимому формализованных полей в записях ChEMBL, а также с помощью регулярных выражений при обработке неформализованных полей.

Аминокислотные последовательности белков человека были извлечены из БД UniProt с помощью следующих фильтров (рис. 9):

- ❖ Статус записи о белке – Reviewed – “Yes”. Записи наиболее достоверны, поскольку составлены под наблюдением эксперта.
- ❖ Биологический таксон – Taxonomy[OC] – “Homo sapiens (Human/Man) [9606]”.
- ❖ Существование белка – Protein Existence[PE] – “Evidence at protein level”. Существование белка доказано на белковом уровне.

Извлекались идентификаторы, канонические аминокислотные последовательности, представляющие референсные сплайс-варианты и сведения о доменном составе. Последнее использовалось при группировке белков по семействам.

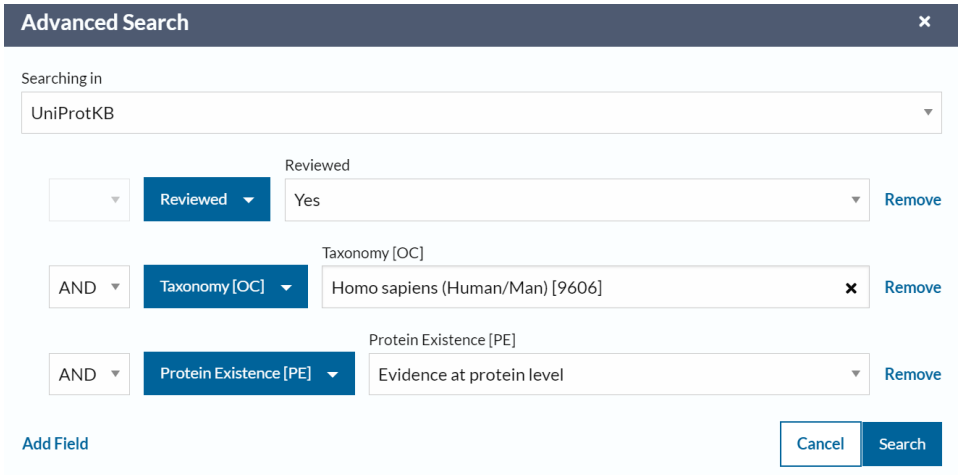

Рисунок 9. Фильтры, используемые при извлечении данных из БД UniProt.

Идентификаторы белков использовались в запросах к БД ChEMBL, которые составлялись в формате MySQL-запроса.

Мы обращались к таблицам по общей характеристике мишеней «Target dictionary», для того чтобы уточнить состав описываемых объектов. Во многих табличных записях под мишенью обозначен комплекс или клеточный лизат, включающий данный белок. Поэтому мы отбирали записи, используя фильтр «Target_type» = «Single protein». В результате был получен список внутренних идентификаторов для мишеней, которые представляли единичные белки.

Запросы по ссылкам на выявленные идентификаторы мишеней позволили извлекать данные о методах, условиях и результатах соответствующих экспериментальных исследований из таблицы «ASSAYS» и «ACTIVITYES» и отбирать нужные записи, устанавливая следующие значения в просматриваемых полях (табл. 1):

Таблица 1. Фильтры, применяемые для отбора данных из БД ChEMBL.

Название поля	**Оператор**	**Значение**	**Характеристика**
ASSAY_TYPE	=	B	Данное значение характеризует биохимические тесты
RELATIONSHIP_TYPE	=	9	Точность обозначения мишени с учетом материала, использованном при исследовании
DATA_VALIDITY_COMMENT	NOT IN*	Outside typical range, Nonstandard unit for type, Potential missing data, Potential transcription error, Potential author error, Author confirmed error	Комментарий, отражающий корректность указанных значений
STANDARD_RELATION	IN**	>, <, =, =>, =<	Показатели связывания должны быть только таких типов. Показатели типа «<<» или «~» являются крайне недостоверными для анализа
STANDARD_VALUE	=	nM	Стандартизованные значения показателя аффинности для исследованных пар
STANDARD_TYPE	IN**	Ki	Тип показателя аффинности должен присутствовать в заданном списке

*NOT IN – значение в данном поле не должно совпадать ни с одним значением из списка.

**IN – значение в данном поле должны соответствовать только указанным в списке.

Отсутствие сведений в любом из указанных полей было критерием для исключения таких данных.

В ответ на запросы мы получали:

- ссылки на химические структуры лигандов, хранящиеся в таблице «COMPOUND_STRUCTURES». Структуры извлекались в двух форматах в форматах «.Mol» и «SMILES», что позволяло восстанавливать структуру в случае дефектных записей «.Mol»;
- показатели взаимодействия для каждой пары;
- идентификаторы аминокислотных последовательностей в базе знаний UniProt;
- сведения из неформализованного поля «DESCRIPTION» (таблица «ASSAYS»).

Опыт работы с данными ChEMBL привел к осознанию того, что обработка неформализованной информации также необходима. Зачастую формализованные поля не содержат важной информации, например, сведений о мутациях, которые являются безусловным поводом для исключения записи из дальнейшей обработки. Для выявления мутаций, обозначенных по типу G55A (замена глицина на аланин в позиции 55) применялось регулярное выражение:

«*.*[ACDEFGHIKLMNPQRSTVWY]{1}\d{1,4}[ACDEFGHIKLMNPQRSTVWY]{1}.*».

В случае с менее определенными формулировками «mutant» или «mutation» применялось следующее выражение:

«.(mutant|mutation).*»*.

Следует отметить, что мы не встретили в проанализированной литературе по РСМ при работе с БД ChEMBL упоминаний о фильтрации по неформализованным полям [Lenselink и соавт., 2017, Sorgenfrei и соавт., 2018, Zakharov и соавт., 2019].

Разработанный нами прогностический алгоритм основан на бинарной классификации обучающих данных. Разделение данных, взятых из ChEMBL, на два класса осуществлялось на основе выбранного порога аффинности, выраженного в виде *Ki*. Заметим, что при более жестких порогах, мы рискуем получить слишком малое число положительных примеров, недостаточное для того, чтобы обеспечить представительность обучающей выборки. Использовалось два пороговых значения *Ki* – более мягкое 10 μM, которое можно рекомендовать

для начального этапа виртуального скрининга и 1 μM для более строгого отбора на последующих этапах.

В результате фильтрации было получено 525 935 записей. Часть из них содержала одни и те же пары белок-лиганд с различными показателями аффинности. Число уникальных пар составило 443542. Потребовалось разработать правила оценки как взаимодействующих, так и невзаимодействующих пар по нескольким записям.

Часто в записи помимо самого лиганда присутствуют молекулы воды, калия, натрия, хлора, сульфат-ион и другие компоненты, которые не рассматриваются как активные соединения. Таким образом, одна и та же основная структура лиганда может быть представлена в нескольких записях, каждая со своим идентификатором. Разными идентификаторами могут быть обозначены стереоизомеры одного и того соединения. Вместе с тем в методах (Q)SAR принято сводить такие варианты к одному объекту, характеризующему активных компонент [Fourches и соавт., 2010]. Для того чтобы устранить избыточность данных в отношении лигандов и описать их в единой унифицированной форме необходимо было провести т.н. процедуру нормализации [Fourches и соавт., 2010]. Для этого мы применили хемоинформатические модули платформы KNIME (рис 10). Все файлы формата MOL были проверены и объединены в единый SDF-файл. Устранялись проблемы, связанные с форматированием. Для удаления ионов, редактирования связей, исключения неизвестных атомов и связей, кекуляризации использовался модуль «RDKit Structure Normalizer».

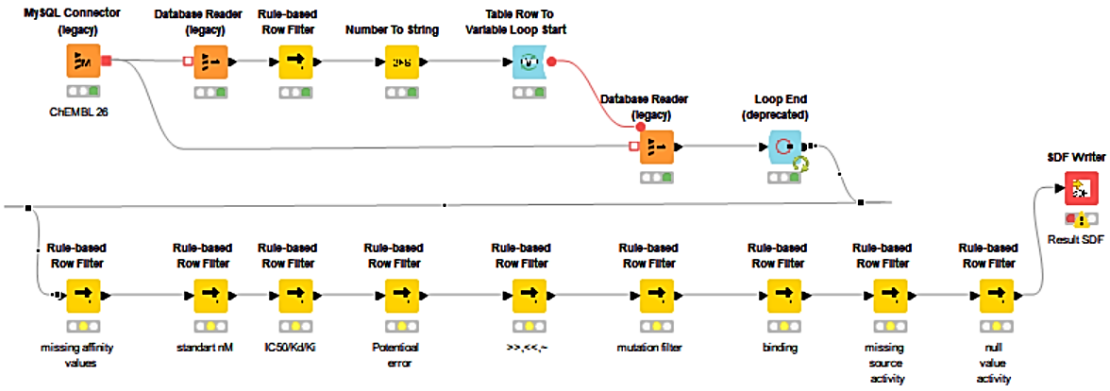

Рисунок 10. Конвейер для извлечения информации из локальной версии ChEMBL под управлением MySQL подготовленный на платформе KNIME.

Стандартизованный набор структур был преобразован в набор MNA-дескрипторов [Filimonov и соавт., 1999] (см. главу 2.3) с помощью утилиты «makeMNA», разработанной в Лаборатории структурно-функционального конструирования лекарств. Для каждой структуры была рассчитана молекулярная масса с помощью модуля «RDkit Structure Calculation». Структуры, которые имели идентичный набор MNA-дескрипторов и одинаковую молекулярную массу, признавались одним соединением и им присваивался обобщенный идентификатор. Все данные о связывании с конкретной мишенью сводились в единый набор, из которого рассчитывались оценки взаимодействия.

Для каждой пары белок-лиганд рассчитывались бинарные оценки взаимодействия путем сравнения показателей взаимодействия с пороговым значением. Оценка взаимодействия принималась равной «1», если установленное значение показателя аффинности не превышало порога, и равной «0» в противном случае.

Нередко данные о связывании для одной и той же пары содержались в нескольких записях. В этих случаях величина, которую сопоставляли с порогом, отбиралась по следующим правилам.

1. Если для пары белок-лиганд найдены как интервальные, так и точные значения показателей связывания, то обрабатываются только точные.
2. Если значений два, то определяется их среднее, а если больше - то медиана.

3. Если для пары найдены только интервальные значения, то во избежание возможных противоречий используется один из трех приемов:
   а) При выявлении записей, содержащих непересекающиеся интервальные значения (например, < 1000 и > 10000), эти записи исключаются из дальнейшей обработки.
   б) Если все величины указаны со знаком «>», то среди них выбирается наименьшее. Если это значение не превышает порог, данные исключаются. Если же установленное значение больше порога оценка взаимодействия приравнивается «0».
   с) Если все величины указаны со знаком «<» меньше, тогда среди них выбиралось наибольшее. Если это значение меньше порога, данные исключаются. Если же установленное значение не превышает порог, то оценка взаимодействия приравнивается «1».

Из БД PDB были извлечены записи, которые представляли трехмерные структуры комплексов белок-лиганд и соответствовали 18% собранных нами обучающих данных. Эти объекты были использованы при валидации метода SPrOS в позиционном режиме. Стоит заметить, что записи трехмерных структур полезны при уточнении обучающих данных в дальнейших проспективных исследованиях. Сведения о трехмерных структурах типа белок-лиганд извлекались по ссылкам на ресурсе PDBbind [Liu и соавт., 2015]. Для сверки идентификаторов лигандов, используемых в PDB, с идентификаторами лигандов в ChEMBL использовался ресурс Ligand Expo [http://ligand-expo.rcsb.org/].

Доменный состав белков-мишеней в соответствии с номенклатурой Pfam уточнялся по записям из БД SIFTS, которые позволяли картировать доменные участки в аминокислотных последовательностях, а также в трехмерных структурах при наличии последних.

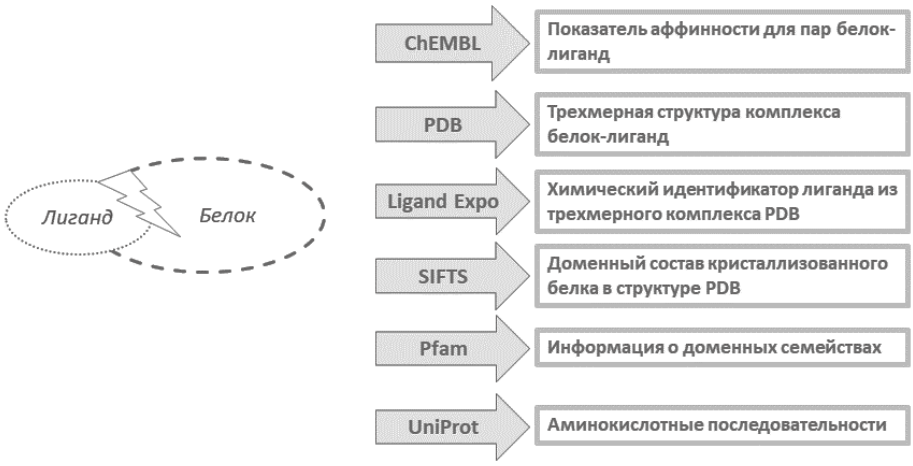

Рисунок 11. Ресурсы, используемые при сборе данных.

Таким образом, реализованная процедура позволила собрать из разных информационных ресурсов сведения об аминокислотных последовательностях белков-мишеней, их доменном составе, расшифрованных трехмерных структурах белков и нормализованных структурах лигандов, а также о показателях взаимодействия, приведенных к единой бинарной форме (рис. 11). Характеристики полученных наборов данных представлены в Таблице 2.

Таблица 2. Характеристики собранных данных о белок-лигандных взаимодействиях.

Группа белков	**Порог аффинности Ki (μM)**	**Размер обучающей выборки белки/лиганды**
GPCR	1	110/4754
	10	112/6411
Протеинкиназы	1	72/277
	10	77/339
Ионные каналы	1	16/15
	10	16/20
Ядерные рецепторы	1	22/55
	10	23/89
Белки из различных семейств	1	252/6443
	10	313/9200

Для проверки эффективности нашего метода в позиционном режиме потребовались наиболее надежные экспериментальные данные, полученные при одинаковых условиях и опубликованные в одних и тех же работах. Мы оценили представительность белков и лигандов в каждом из исследований, указанных в ChEMBL, результаты которых были включены в обучающие данные.

Одна из найденных статей [Karaman и соавт., 2008] представляла исследование протеинкиназ человека (255 доменов), сгруппированных в соответствии с величинами константы диссоциации *Kd*, полученных при воздействии низкомолекулярных соединений *in vitro*. В публикации содержались детальные сведения об использованных препаратах и экспериментальных условиях. Мы включили в обучающие данные результаты, полученные при тестировании отдельных киназных доменов. Оценка точности прогноза производилась путем проецирования предсказанных позиций на трехмерные структуры белков, содержащие класс-образующий лиганд в кармане связывания. Также с этой целью привлекались сведения о мутантных вариантах с измененной аффинностью к ингибиторам. Трехмерные структуры, использованные для оценки точности, представлены в Таблице 3. Пары белок-лиганд отобраны с учетом спектра ингибирования протеинкиназы (подробнее в следующем разделе).

Таблица 3. Выбранные для валидации позиционного режима комплексы белок-лиганд.

Идентификатор UniProt	**Название белка**	**Ингибитор**	**Идентификатор PDB**
P00519	Tyrosine-protein kinase ABL1	Иматиниб	2HYY
Q96GD4	Aurora kinase B	VX-680	4AF3

Таким образом, мы отобрали данные для тестирования нашего подхода в позиционном, а также в прогностическом режиме.

3.3 Подготовка данных из «Stanford HIV database»

Для оценки эффективности метода при оценке взаимодействия близкородственных белков с лигандами использовался ресурс, посвященный лекарственной резистентности белков ВИЧ (вируса иммунодефицита человека) «Stanford HIV database». Эта БД содержит сведения о резистентности мутантных вариантов трех белков ВИЧ к различным ингибиторам, которые, в свою очередь разделялись на четыре группы (табл. 7). Аминокислотные последовательности восстанавливались из множественного выравнивания. В ряде последовательностей отдельные позиции содержали более одной буквы в связи с неоднозначными результатами секвенирования. Такие последовательности исключались из дальнейших исследований в связи с невозможностью точного восстановления исходных последовательностей.

3.4 Оценка точности прогноза в позиционном режиме с использованием модельных аминокислотных последовательностях

Подход к оценке точности предсказания функционально значимых аминокислотных позиций с использованием модельных последовательностей представляется вполне рациональным [Kalinina и соавт., 2004]. Это связано с неполнотой данных о вкладе аминокислотных остатков в распознавании лигандов. Кроме того, в случае естественных белков картина нередко осложняется наличием нескольких групп остатков, отвечающих за разные молекулярные взаимодействия, а также консервативных позиций, отражающих «эволюционный след». Путем внесения класс-специфичных замен в модельные последовательности на четырех участках R1, R2, R3 и R4 (рис. 12), были смоделированы три искусственных подсемейства по семь последовательностей каждое. Часть замен находилась в различных колонках, что наблюдается и при сопоставлении естественных белков-гомологов, когда положение позиций, консервативных для отдельных подгрупп может не совпадать при выравнивании (участки R1m2 и R4m2, рис. 12).

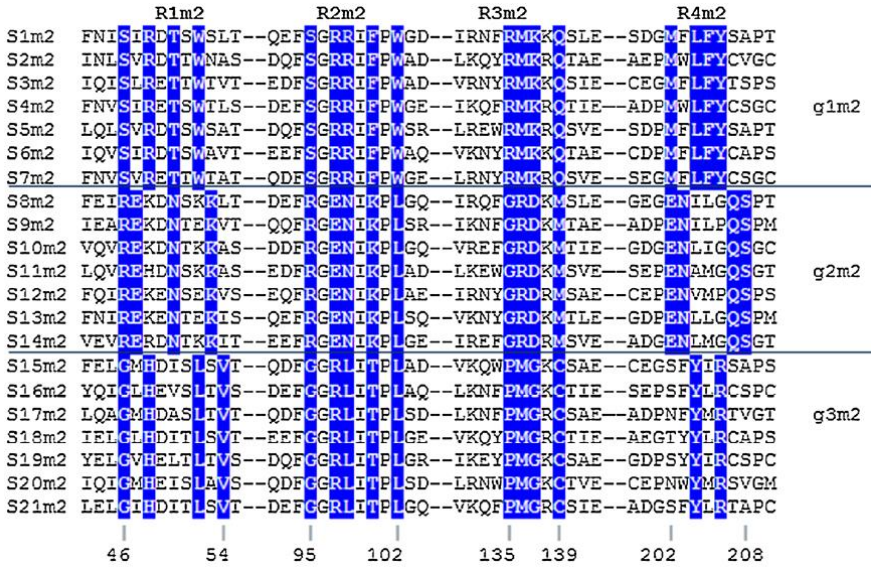

Рисунок 12. Участки искусственных последовательностей R1m2, R2m2, R3m2 и R4m2 с «мутациями» специфичными для подгрупп g1m2, g2m2 и g3m2 выделены цветом.

При испытании подхода в позиционном режиме тестовая последовательность исключалась из обучающей выборки, на нее проецировались все остальные последовательности, как семейства, так и его дополнения (прочих «белков»). Результаты, полученные для одной из смоделированных последовательностей (S1m2), представлены на рисунке 13. Как видно, все замены, характерные для группы g1m2, идентифицированы со статистически значимыми оценками (*p*-значение < 0,001) (рис. 13). *P*-значения для остальных остатков находятся на фоновом уровне. Важно отметить, что значимые оценки были получены как для участков с несмещенными (R2m2, R3m2), так и смещенными (R1m2, R4m2) заменами. Та же процедура была проведена для всех белков искусственного семейства.

Общая оценка точности прогноза была проведена на основе позиционных оценок $E_{pa,}$ полученных для всех как специфичных, так и неспецифичных позиций всех последовательностей. Величина *IAP* составила более 0.99. Небольшое отклонение от максимально возможного значения *IAP* (1,0) можно объяснить сложной конфигурацией замен в областях R1m2 и R4m2. Таким образом, была продемонстрирована принципиальная способность нашего метода распознавать позиции, специфичные для подгрупп последовательностей.

Рисунок 13. Результат определения специфичных аминокислотных остатков для последовательности S1m2. *P*-значения отложены в обратной логарфимической шкале.

3.5 Оценка точности прогноза в позиционном режиме на примере протеинкиназ их ингибиторов

Для тестирования позиционного режима на естественных последовательностях было выбрано семейство протеинкиназ человека. При этом мы использовали весьма надежный набор данных [Karaman и соавт., 2008], который представлял полную матрицу 317 ферментов и 38 ингибиторов, заполненную экспериментально определенным показателям связывания (Kd), для каждой пары. Для испытания были отобраны 255 белков, которые представляли собой отдельные протеинкиназые домены. Это позволило точнее определить характер взаимодействия, поскольку многие белки этого семейства являются многодоменными.

Основным критерием при выборе тестовых протеинкиназ было наличие разрешенного трехмерного комплекса фермента с класс-образующим лигандом. Это позволило остановиться на двух таких комплексах (табл. 3, рис. 14).

Важно отметить, что исследуемые ингибиторы проявляют различные спектры активности в сопоставлении с филогенетическими отношениями в семействе протеинкиназ (рис. 14).

ABL1 – Иматиниб – 2HYY ***AurkB – VX-680 – 4AF3***

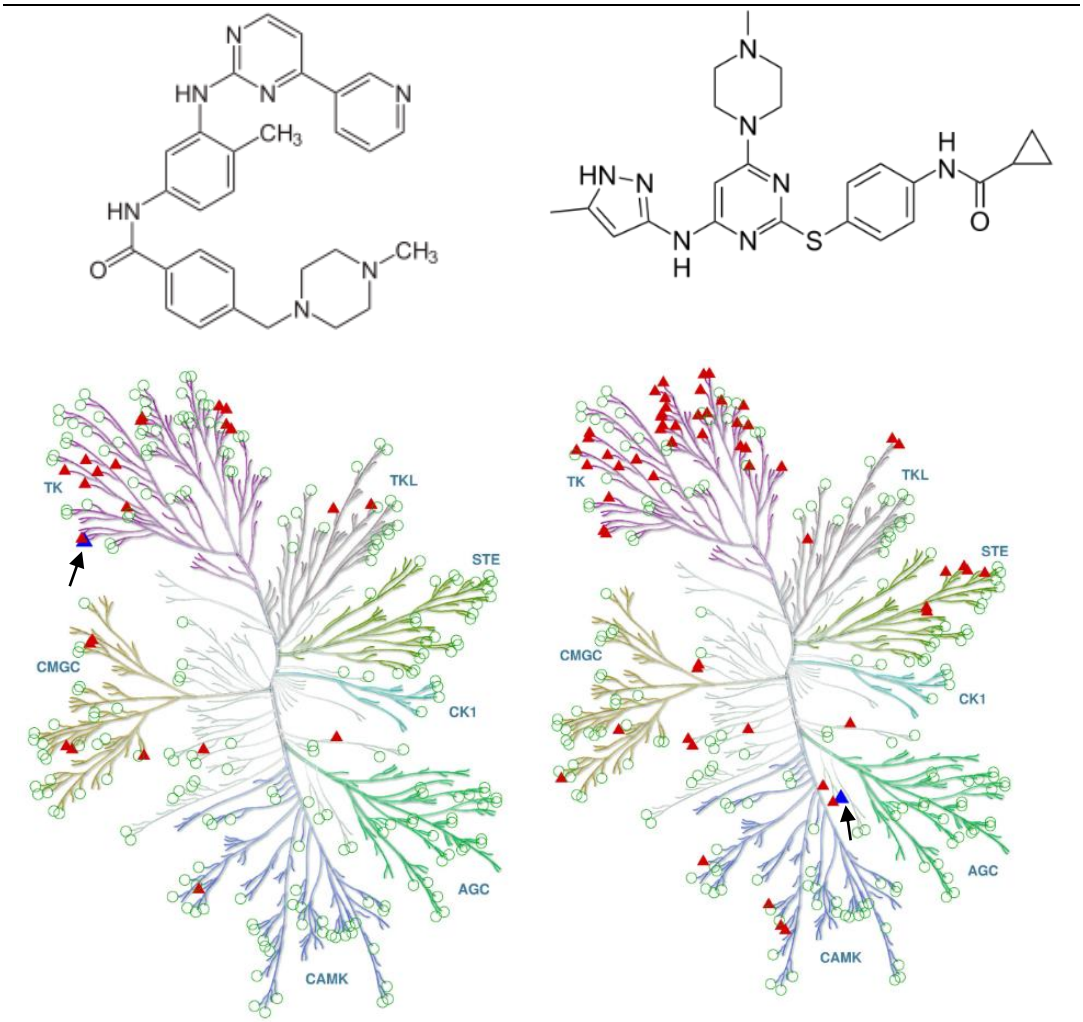

Рисунок 14. Спектр ингибирования лигандов, выбранных для тестирования. Вверху – названия белков-мишеней, название ингибитора и идентификатор комплекса PDB. Отражены структуры низкомолекулярных ингибиторов. На филогенетические деревья нанесен спектр ингибирования при пороге Kd = 1 мкм. Красными треугольниками показаны киназы, ингибируемые при значениях Kd ниже порога. Зеленые кружки соответствуют более высоким значениям Kd. Синими треугольниками обозначены тестируемые белки-мишени (дополнительно отмечены стрелочками).

Так, Иматиниб активен по отношению к тирозинкиназам (кластер ТК). Киназа ABL1, образующая тестируемую пару с Иматинибом является представителем данной группы. Ингибитор VX-680 также преимущественно воздействует на объекты кластера ТК, однако активен и против серин-треониновых протеинкиназ. Он характеризуется более широким спектром ингибирования по сравнению с Иматинибом. Для данного лигнада мы выбрали

киназу AurkB, которая отстоит достаточно далеко от кластера ТК на филогенетическом дереве.

При предсказании аминокислотных позиций ABL1, специфичных для взаимодействия с Иматинибом, статистически значимые оценки получили 29 остатков (*p*-значение < 0,001). При анализе трехмерного комплекса 2HYY (Иматиниб-ABL1) было показано, что пять из предсказанных остатков локализованы в области связывания лиганда и непосредственно контактируют с ингибитором (рис. 15). Так, T315 формирует водородную связь с ингибитором, благодаря гидроксильной группе. Замена T315I приводит к снижению аффинности в более чем 100 раз [Karaman и соавт., 2008]. Остальные четыре остатка A269, V270, K271 и I313 участвуют в гидрофобном взаимодействии с молекулой ингибитора.

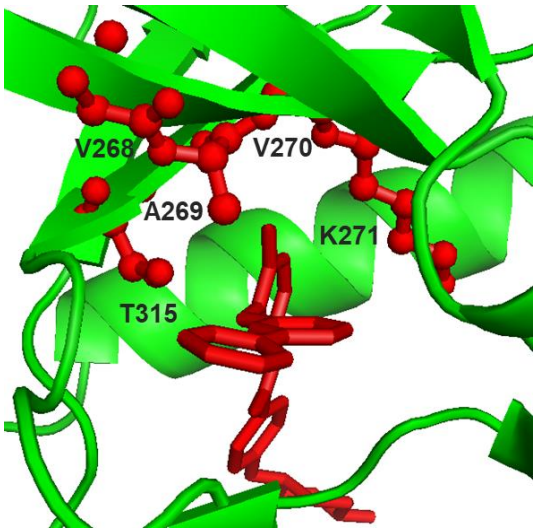

Рисунок 15. Область связывания ингибитора Иматиниб с протеинкиназой ABL1. Красным отмечена молекула ингибитора. Красными сферами отмечены предсказанные аминокислотные остатки.

Предсказанные остатки вне области непосредственного контакта с лигандом также могут влиять на взаимодействие с ним. Шесть остатков из N-доли могут опосредованно влиять на конформацию АТФ-кармана в связи с близкой локализацией к нему.

Существенная часть остатков, получивших значимые оценки локализованы удаленно. Большинство из них расположено в области связывания

фосфорилируемого белка-субстрата. Вероятно, высокие оценки были получены из-за того, что обучающая выборка содержала ряд близких гомологов ABL1, которые имеют близкий репертуар белков-партнеров. Для проверки этого предположения мы исключили из обучающей выборки пять наиболее сходных с ABL1 белков, при этом попарная идентичность оставшихся аминокислотных последовательностей не превышала 48% (рис. 16).

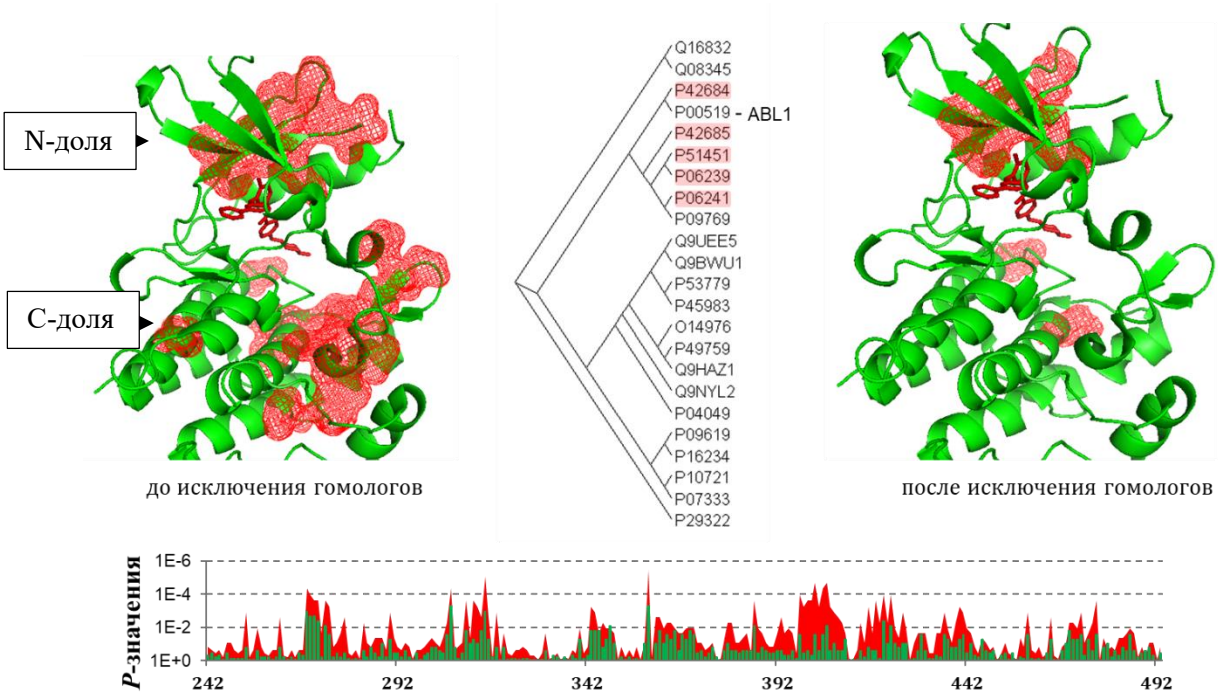

Рисунок 16. Результат прогноза специфичных аминокислотных остатков в последовательности протеинкиназы ABL1 к ингибитору Иматиниб. Сетью обозначены аминокислотные остатки, получившие статистически значимые оценки. Красным на филогенетическом дереве отмечены исключенные последовательности. *P*-значения, рассчитанные до и после исключения ближайших гомологов показаны красной и зеленой областями, соответственно.

В результате количество предсказанных остатков снизилось до восьми. Четыре остатка из области лиганд-связывающего кармана T315, A269, V270 и K271 были предсказаны до исключения близких гомологов. Наибольшую оценку получил остаток N358, который вероятно, обеспечивает конформационную подвижность активационной петли протеинкиназы. Важно отметить, что в структуре ABL1-киназы со свободным АТФ-карманом активационная петля лежит плотно на белковой глобуле (рис. 17). В такой нативной структуре образуются две водородные связи между N358 и M388. Связывание же с

Иматинибом приводит к разрыву данных связей и смещению активационной петли. Более прочная фиксация активационной петли, обусловленная другим остатком в данной позиции, может препятствовать встраиванию молекулы ингибитора. Стоит отметить, что остаток M388 получил достаточно низкую оценку. Поскольку связи между N358 и M388 формируются за счет радикала первого остатка и основной цепи второго, радикал M388 может существенно не влиять на специфичность взаимодействия.

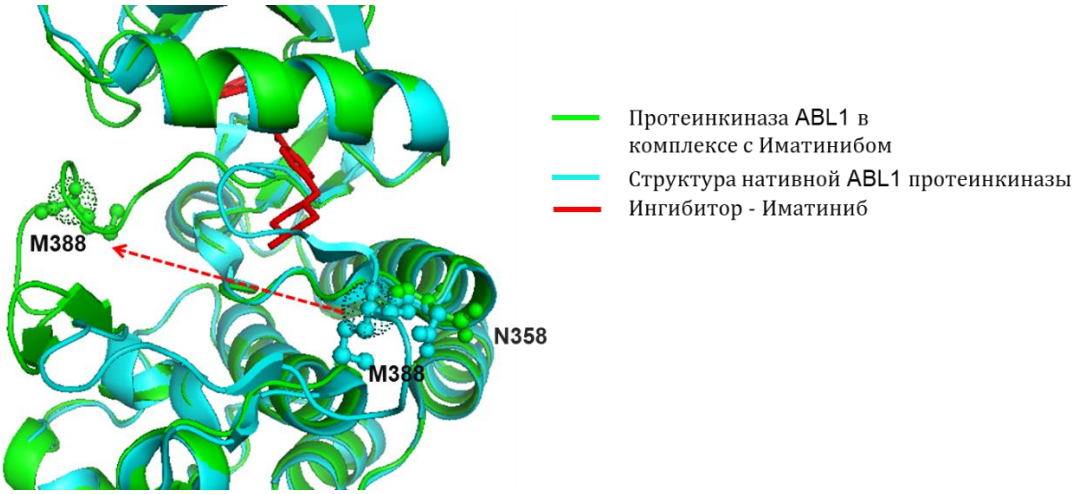

Рисунок 17. Трехмерное выравнивание комплекса ABL1-Иматиниб и нативной ABL1 протеинкиназы (2HZ4) отражает конформационное смещение, которое происходит при связывании с иматинибом в сравнении с нативной структурой. Стрелкой обозначено изменение положения M388, Cα-атом M388 обозначен полупрозрачной сферой.

При предсказании позиций, критичных для связывания протеинкиназы AurkB и ингибитора VX-680 мы выявили шесть остатков со значимыми оценками. Пять из них (L83, Y156, K215, I216, A217) локализованы в области контакта белок-ингибитор (рис. 18), а у шестого (S222) не установлены координаты. Более определенные результаты в сравнении с ранее описанной парой можно объяснить относительно небольшим числом близких гомологов AurkB в обучающей выборке, что снижает влияние филогенетических отношений.

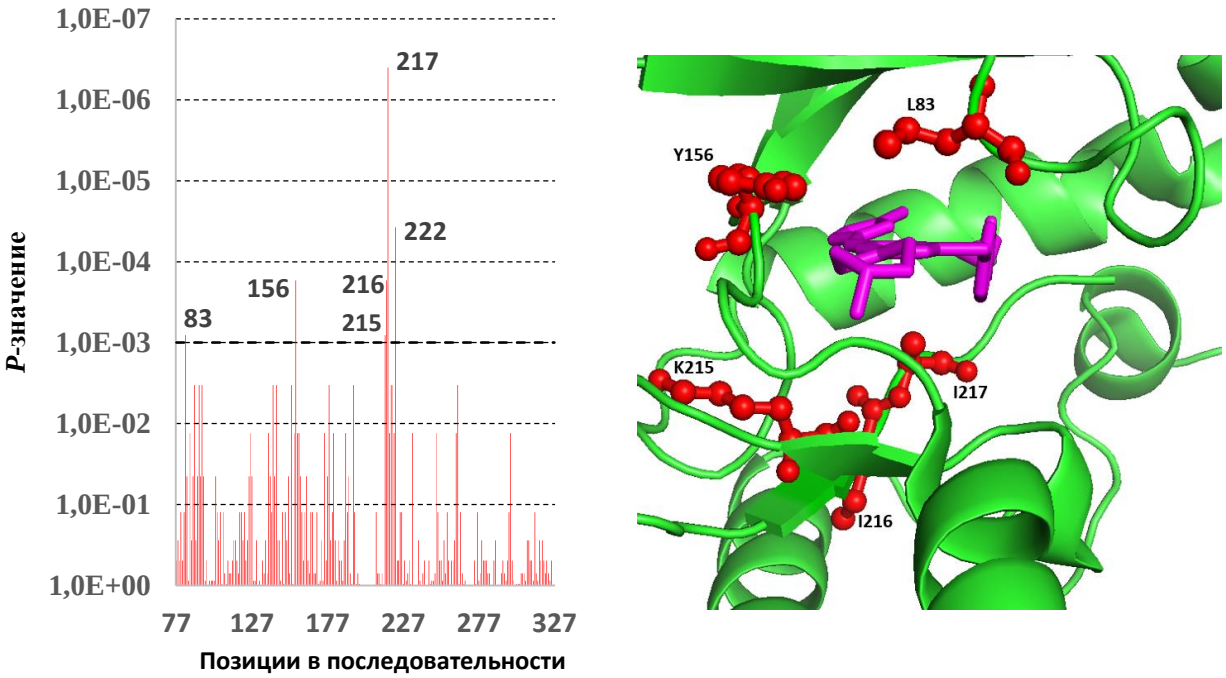

Рисунок 18. Статистические оценки, полученные для остатков протеинкиназы AurkB. Проекция предсказанных остатков на трехмерный комплекс AurkB с ингибитором VX-680. Ингибитор отмечен фиолетовым.

В результате данного исследования было показано, что в разработанном подходе аминокислотные остатки, которые могут влиять на связывание с лигандами, получают наиболее значимые оценки. Часть из предсказанных остатков контактирует с ингибиторами, для других – удается найти подтверждение в литературе, а также путем сопоставления трехмерных структур. В связи с этим, позиционные оценки могут использоваться при расчете интегральной вероятностной оценки взаимодействий для пар белок-лиганд. Процедура фильтрации близких гомологов белка-мишени (исключения из обучающей выборки) может быть полезна для того, чтобы точнее локализовать область связывания лиганда или уточнить его положение.

Важно отметить, что позиционный режим позволяет наглядно оценить идею идентификации локального соответствия между белками и лигандами, при этом получая интерпритируемые результаты.

3.6 Оценка точности прогноза в прогностическом режиме по второму сценарию

Данные из статьи [Karaman и соавт., 2008] были также использованы для оценки эффективности нашего подхода в прогностическом режиме по второму сценарию, то есть без учета структур лигандов (в связи с малым количеством лигандов). Величина *IAP* составила 0,80, что указывало на достаточно высокую точность.

Обучающая выборка «Gold Standard» [Yamanishi и соавт., 2008] использована в большом числе независимых исследований при оценке точности различных прогностических методов [Xia и соавт., 2010, Van Laarhoven и соавт., 2013; Chen и соавт., 2013, Pahikkala и соавт., 2015; Zheng и соавт., 2013; Liu и соавт., 2016; Shi и соавт., 2018]. Поэтому мы применили эти данные для сопоставления нашего метода с другими (табл. 4) с учетом того, что составители «Gold Standard» предложили бинарное разбиение набора на взаимодействующие и невзаимодействующие пары белок-лиганд. Сравнение было проведено только по второму сценарию в связи с тем, что ряд белков имел малое количество лигандов (1-2), что не позволяло провести тестирование по третьему сценарию.

Важно отметить, что в указанных в Таблице 4 подходах применялись различные стратегии валидации. Например, методы «TMF» и «NRLMF» валидировались с помощью десятикратного разбиения обучающей выборки, т.е. 1/10 часть данных выступала в качестве теста, 9/10 в качестве обучающих данных [Liu и соавт., 2016; Shi и соавт., 2018]. Авторы метода «CMF» разделяют обучающую выборку на пять частей, так 1/5 часть данных используется в качестве тестового набора, остальные 4/5 – в качестве обучающих данных [Chen и соавтр., 2013]. Таким образом, провести полностью объективное сравнение не представляется возможным. Тем не менее, такое тестирование позволяет определить принципиальную возможность осуществлять прогноз белок лигандных взаимодействий с учетом уже разработанных подходов.

Таблица 4. Значения ROC AUC*, полученные при использовании разных методов.

Белки-мишени	**Название метода**	
	F	**SPrOS**
Ферменты		
Ионные каналы		
Ядерные рецепторы		

* Оценивалась точность прогноза с помощью величины *IAP*, рассчитанной с использованием процедуры скользящего контроля с исключением по одному, которая численно равна величине ROC AUC, использованной другими авторами. Оценки точности других методов получены из соответствующих публикаций.

Видно, что оценки точности прогноза взаимодействия ионных каналов с их лигандами достигали высоких значений при тестировании всех методов – от 0,91 до 0,97. Наш метод также показал для этой группы высокую оценку равную 0,92.

В случае с ферментами оценки, полученные с помощью разных методов, распределились в более широком диапазоне от 0,88 до 0,98. Это можно объяснить тем, что группа ферментов собрана без учета филогенетического родства мишеней, которые представляет белки из разных семейств. Тем не менее, наш метод продемонстрировал высокую точность - 0,94. Для семейства GPCR ряд методов продемонстрировал весьма низкие оценки, значения которых опускались до 0,77. Наш подход показал высокую оценку равную 0,94. Наиболее противоречивые результаты получены для ядерных рецепторов; так, для четырех методов точность оценивалась низкими значениями (0,66 – 0,74), при том, что другие показали приемлемую точность. Наш метод показал наивысшую оценку – 0,99.

Таким образом, метод SPrOS позволяет предсказывать взаимодействие белок-лиганд с высокой точностью от 0,92 до 0,99 при обучении на разных типах

данных. При тестировании подхода по второму сценарию были достигнуты результаты, сопоставимые или превосходящие по точности результаты других авторов.

3.7 Оценка области применимости похода при различных сценариях прогностического режима

Мы провели испытания нашего подхода на обучающих наборах данных, различающихся по количеству взаимодействующих пар белок-лиганд и доменному составу белков-мишеней.

3.7.1 Обучающие данные без группировки белков мишеней

Задача такого рода возникает достаточно часто на первых этапах исследования. Разнородность обучающих данных, как по мишеням, так и по лигандам в данном наборе максимальна, что может сказаться на эффективности прогноза. В то же время, работа с подобными выборками позволяют более полно оценить применимость метода в различных областях исследования. Белки-мишени, включенные нами в этот обучающий набор, содержат в совокупности более 60 доменных типов, каждый из которых характеризуется одним или несколькими участками связывания. Эти структурные особенности обуславливают то, к каким классам химических соединений должны относиться установленные или возможные лиганды.

Первый сценарий прогностического режима (рис. 4, рис. 8, глава 2.5), подразумевает прогноз белков-мишеней без учета их аминокислотных последовательностей на основе структурного описания лигандов с помощью программы PASS. Тестирование показало высокую точность прогноза (0,9) при обоих порогах – в 1 и в 10 μM (табл. 5). Это свидетельствовало как об эффективности применённого метода, так и о существенном структурном сходстве лигандов, специфичных к одним и тем же мишеням. Последнее соображение послужило еще одним основанием для использования оценок

прогноза PASS в качестве коэффициентов лигандной специфичности при работе по третьему сценарию.

Второй сценарий предусматривает прогноз лигандной специфичности белков на основе их аминокислотных последовательностей. Идентификаторы лигандов используются как класс-образующие признаки без учета их структур. Тестирование по второму сценарию также показало высокие оценки при обоих порогах активности – 0.98.

Таблица 5. Тестирование на несгруппированных мишенях.

Порог аффинности Ki (μM)	***IAP***			**Размер обучающей выборки**
	1й сценарий	**2й сценарий**	**3й сценарий**	**белки/лиганды**
1	0.98	0.98	0.91	252/6443
10	0.98	0.98	0.86	313/9200

При реализации третьего сценария используется описание, как мишеней, так и лигандов. При тестировании обоих порогов получено заметное снижение точности – на 0,07 при пороге в 1 μM и на 0,12 при пороге в 10 μM. Такое снижение вероятно, связано со значительным расширением пространства признаков вследствие размытия границ классов за счет использования нечетких коэффициентов лигандной специфичности. Тем не менее, во всех сценариях прогноза получены высокие оценки, что подтверждает применимость метода на наиболее разнородных наборах данных. Модели, обученные на таких данных, могут быть полезны на ранних этапах исследования, когда необходимо установить, например, лиганды для белка-мишени с неустановленным спектром мишеней, а также для прогноза мишеней биологически активного лиганда.

3.7.2 Обучающие данные с разбиением мишеней на классы белков

Группы белков, сходные по структурным и функциональным свойствам, часто исследуются при поиске лекарственных мишеней и разработке лекарств. На данном этапе мы по отдельности использовали для тестирования наборы данных, сформированные на основе классов перспективных лекарственных мишеней (табл. 6).

Таблица 6. Тестирование на мишенях разных классов.

Группа белков	**Порог активности (μМоль)**	**IAP**		
		1й сценарий	**2й сценарий**	**3й сценарий**
GPCR	1	0,99	0,98	0,90
	10	0,98	0,98	0,92
Протеинкиназы	1	0,98	0,96	0,89
	10	0,98	0,94	0,86
Ионные каналы	1	0,99	0,98	0,96
	10	0,99	0,97	0,96
Ядерные рецепторы	1	0,99	0,99	0,98
	10	0,99	0,99	0,99

При тестировании по первому сценарию, то есть при предсказании мишеней для лигандов (рис. 4, рис. 8) была показана высокая точность с величинами IAP не менее 0.98 при обоих порогах активности. При тестировании по второму сценарию оценки точности также достигали высоких значений. Некоторое снижение точности 0,94 (1 μM) и 0,96 (10 μM) отмечено при прогнозе взаимодействия протеинкиназ с их ингибиторами.

При тестировании по третьему сценарию отмечено существенное снижение оценок для всех групп - также и в случае не сгруппированных мишеней. Больше всего точность снизилась при тестировании на данных, относящихся к протеинкиназам, упав до 0,86 при пороге в 10 μМоль. При валидации с двумя группах мишеней GPCR и ионными каналами значения IAP снизились до 0,9 и 0,96, соответственно. При испытаниях на данных с ядерными рецепторами оценка точности превышала 0,98 при всех сценариях. Несмотря на снижение точности в

третьем сценарии, она оставалась на высоком уровне. Таким образом, можно рекомендовать этот метод для проспективных исследований.

Степень снижения точности прогноза при третьем сценарии зависит от характера обучающих данных. Существенное снижение показано для более представительных наборов, относящихся к GPCR и протеинкиназам, что может объясняться их большим разнообразием. Наиболее выражено падение оценок точности в случае с протеинкиназами. Различные ингибиторы ферментов этого семейства связываются со структурно консервативным АТФ-карманом, проявляя, однако «парадоксальную» специфичность к отдельным представителям семейства [Thaimattam и соавт., 2007]. По-видимому, структурные характеристики лигандов, которые могут быть связаны с избирательной аффинностью, частично нивелируются при использовании коэффициентов, рассчитанных по сходству этих лигандов.

Тенденция к снижению точности при комбинированном использовании дескрипторов лиганда и белка-мишени показана и другими авторами как для протеинкиназ [Sorgenfrei и соавт., 2018], так и для белков других семейств [Shi и соавт., 2018]. Необходимо упомянуть, что в указанных работах валидация осуществлялась при одновременном исключении белка и лиганда из обучающей выборки. Однако такой протеохемометрический подход, реализованный в третьем сценарии, востребован в связи с неполнотой обучающих данных. Подобная нехватка информации типична для большинства задач, решаемых при поиске биологически активных соединений. Использование выборок, представляющих полные матрицы взаимодействия (рис. 4) обеспечит более точный прогноз, но чрезвычайно сузит область применимости метода.

3.7.3 Оценка применимости метода в случае близко гомологичных белков-мишеней на примере белков ВИЧ.

Стендфорская базы данных (Standford HIV database) предоставляет данные по эффективности ингибиторов в связи с наличием мутаций в трех белках «дикого

типа» - протеазы, обратной транкскриптазы и интегразы. Мутанты каждого белка рассматривались как близкие гомологи, содержащие небольшое количество замен и составляли вместе с ингибиторами отдельную обучающую выборку. Последовательности обратной транскриптазы были включены в две выборки в зависимости от типа ингибиторов – нуклеозидных и ненуклеозидных.

Все классы ингибиторов были представлены небольшим числом соединений – восемью для протеазы, шестью нуклеозидными и четырьмя ненуклеозидными для обратной транскриптазы и четырьмя для интегразы. Это не позволило провести испытание по первому и третьему сценариям.

В отличие от предыдущих наборов последовательностей в данном случае специфичность взаимодействия зависит от изолированных точечных мутаций. Поэтому вы провели тестирование с двумя значениями длины сопоставляемых фрагментов $\boldsymbol{F}$ – 7 и 30.

Каждая из четырех выборок представляла полную матрицу взаимодействия, так что все белки разбивались в отношении каждого лиганда на классы взаимодействующих и невзаимодействующих в соответствии с выбранным порогом отношения резистентности RR (см. главу 3.3). Валидация проводилась при шести значениях порога (1,5, 2, 4, 6, 8, 10) и в качестве результата рассматривалась наибольшая из величин, полученных при шести испытаниях и двух значениях $\boldsymbol{F}$ (табл. 7).

Тестирование по второму сценарию показало высокую точность прогноза для большинства пар белок-ингибитор (табл. 7). Так, в случае протеазы оценка точности варьировала от 0,93 (Типранавир) до 0,98 (Лопинавира). Оценки точности для обратной транскриптазы и нуклеозидных ингибиторов располагались в диапазоне от 0,85 до 0,99, а для пар обратной транскриптазы и ненуклеозидных ингибиторов точность оказалась ниже. Достаточно низкие значения IAP были получены для Рилпивирина и Этравирина - 0,713 и 0,765, соответственно. В то же время для двух других ингибиторов данного класса Эфавиренза и Невирапина точность прогноза была достаточно высокой - 0,874 и 0,899. Высокую точность удалось обеспечить не для всех ингибиторов интегразы.

Тем не менее, для 18 из 22 исследованных ингибиторов были получены высокие оценки точности, в большинстве случаев превышающие 0,9, что указывает на целесообразность проведения проспективных исследований с целью оценки резистентности новых штаммов.

Таблица 7. Тестирование на белках ВИЧ и их ингибиторах по второму сценарию.

Мишень ингибитора	Название препарата	Количество вирусных вариантов	Максимальная оценка точности	Порог (RR)
Протеаза	*Атазанавир*	451	0,952	10
	Дарунавир	245	0,978	4; 10
	Фосампренавир	720	0,954	2
	Индинавир	750	0,973	4
	Лопинавир	596	0,983	2
	Нелфинавир	771	0,955	4
	Саквинавир	759	0,951	2
	Типранавир	302	0,934	10
Обратная транскриптаза (нуклеозидные ингибиторы)	*Ламивудин*	445	0,953	10
	Абакавир	443	0,95	1,5
	Зидовудин	443	0,926	2
	Ставудин	445	0,944	6
	Диданозин	445	0,999	10
	Тенофовир	350	0,845	4
Обратная транскриптаза (ненуклеозидные ингибиторы)	*Эфавиренз*	512	0,874	8
	Этравирин	135	0,765	4
	Невирапин	517	0,899	6; 8
	Рилпивирин	74	0,713	8
Интеграза	*Биктегравир*	101	0,783	4
	Долутегравир	155	0,72	2
	Элвитегравир	270	0,857	1,5
	Ралтегравир	287	0,862	2

Стоит обратить внимание на то, что для большинства ингибиторов максимальная точность достигается на невысоких порогах резистентности в 1,5, 2 и 4, что может обеспечить раннее выявление новых резистентных штаммов вируса.

3.8 Точность прогноза при разных значениях параметра F

В молекулярных взаимодействиях важную играют роль кооперативные эффекты аминокислотных остатков белковой молекулы. Для оценки совместного влияния остатков, расположенных в последовательности на различных расстояниях, мы провели испытания при разных значениях параметра *F*, который определяет длину сравниваемых фрагментов (рис. 19).

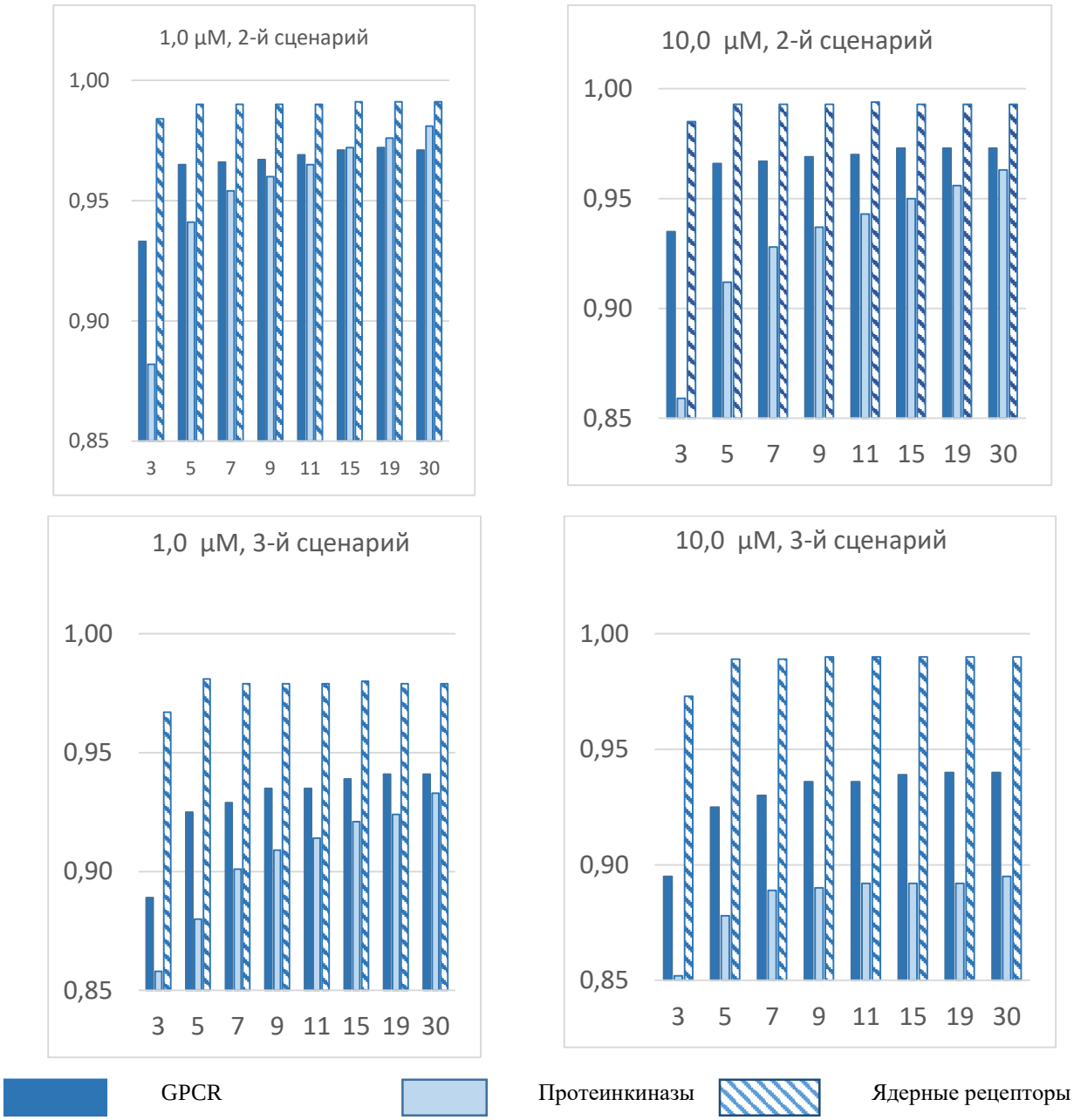

Рисунок 19. Значения *IAP* (ось Y) при различных значениях параметра *F* (ось X). Валидация проводилась при скользящем контроле с исключением по одному.

Мы протестировали наш метод по второму и третьему сценарию прогностического режима при значениях *F* от 3 до 30 остатков. Валидация была проведена на трех ранее вышеупомянутых выборках, представляющих GPCR, протеинкиназы, ядерные рецепторы и их лиганды. Было показано, что точность прогноза возрастает с увеличением *F* как при втором, так и при третьем сценарии. Это, по-видимому, связано с влиянием отдаленных межпозиционных зависимостей на специфичность белка к лигнаду. Хотя применение нечетких коэффициентов взаимодействия в третьем сценарии понизило точность прогноза, полученные оценки оставались достаточно высокими. Это еще раз подтвердило применимость нашего подхода в ситуации, моделируемой 3-м сценарием. Наибольшие различия, связанные с величиной порога, отмечены при тестировании на протеинкиназах, что может быть связано с вышеописанными особенностями АТФ-кармана, влияющими на аффинность лиганда к белку [Thaimattam и соавт., 2007].

3.9 Веб-сервис для прогноза белок-лигандных взаимодействий в трех сценариях

Существующие на данный момент методы протеохемометрики, как правило, недоступны для широкого круга исследователей. Предоставляемый в открытом доступе инструментарий обычно рассчитан на специалистов, которые владеют методами машинного обучения и способны самостоятельно разрабатывать прогностические модели. В качестве примера можно указать библиотеку языка программирования R «camb», которая включает в себя инструменты для генерации дескрипторов белков и лигандов, построения моделей, визуализации и др. [Murrell и соавт., 2015].

Сервис, предоставляющий уже обученные прогностические модели, позволяет оперативно определять круг мишеней и лигандов, наиболее перспективных для дальнейшей разработки с применением экспериментальных методик. Созданный нами веб-сервис «http://way2drug.com/proteochemometrics/»

позволяет прогнозировать белок-лигандные взаимодействия в наиболее распространенных сценариях, которые возникают при поиске новых лекарственных средств (рис. 20). Сервис обеспечивает предсказание белков мишеней тестируемого лиганда по первому сценарию, принимая на вход структурную формулу низкомолекулярного соединения. При предсказании лигандов для тестируемой мишени по второму сценарию на вход подается только аминокислотная последовательность. Процедура, реализующая третий сценарий (протеохемометрика) требует в качестве входных данных как структурную формулу лиганда, так и аминокислотную последовательность.

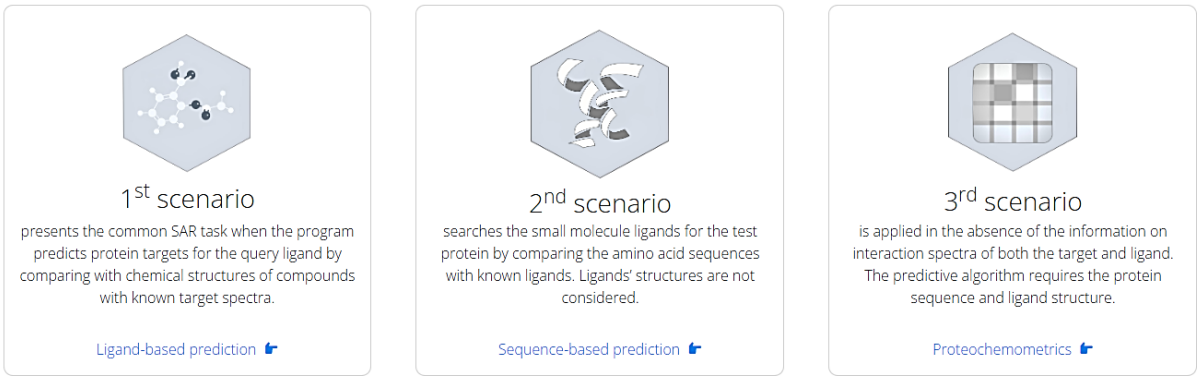

Рисунок 20. Сценарии прогноза, реализованные на веб-сервисе «http://way2drug.com/proteochemometrics/».

Рассмотрим использование веб-сервиса на примере третьего сценария (рис 21). При вводе данных необходимо соблюдать следующие условия:

- Лиганд должен содержать как минимум три углеродных атома.
- Структура должна быть электронейтральной, на вход подается только однокомпонентная структура, без изолированных атомов.
- Молекулярная масса лиганда не должно превышать 1250 Да.
- Аминокислотная последовательность вводится в формате «FASTA» [Lipman and Pearson, 1985] и не должна содержать иных символов кроме обозначений 20 канонических остатков.
- Аминокислотная последовательность должна содержать не менее 30 букв.

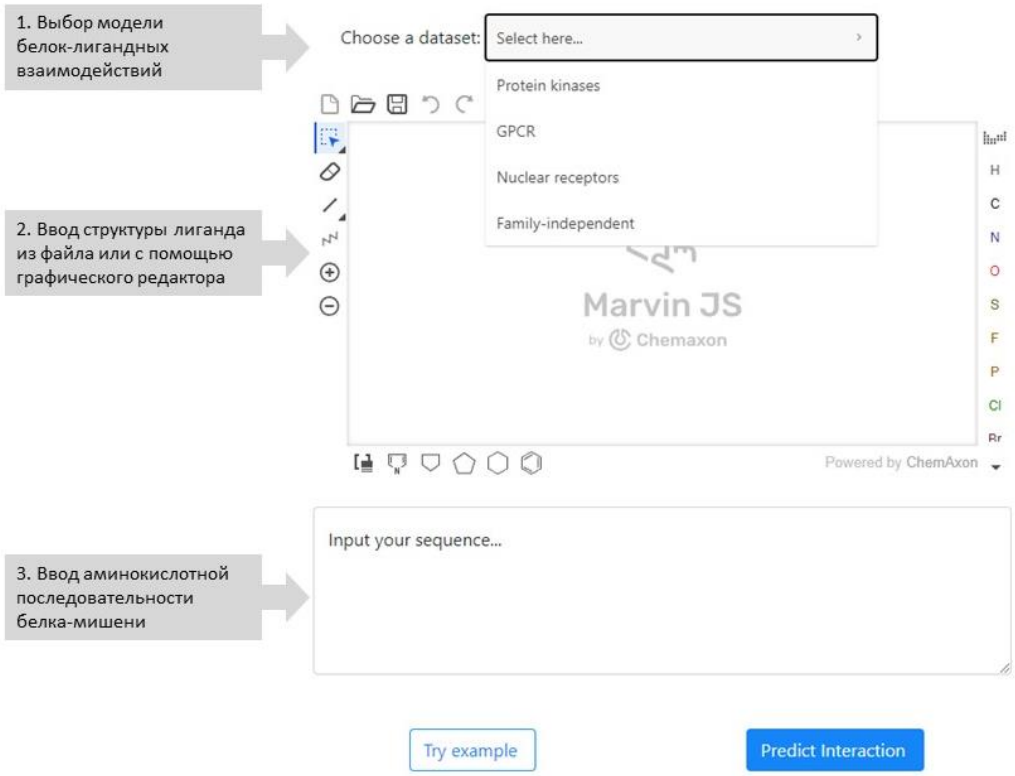

Рисунок 21. Интерфейс веб-сервиса с полями для ввода данных.

В действующей версии веб-сервиса возможен прогноз на основе четырех обучающих выборок (моделей взаимодействий), включающих данные по следующим группам белков-мишеней: протеинкиназы, ядерные рецепторы, ионные каналы, а также выборка, которая составлена по сведениям о белок-лигандных взаимодействиях для белков из различных семейств (рис. 21, рис. 22).

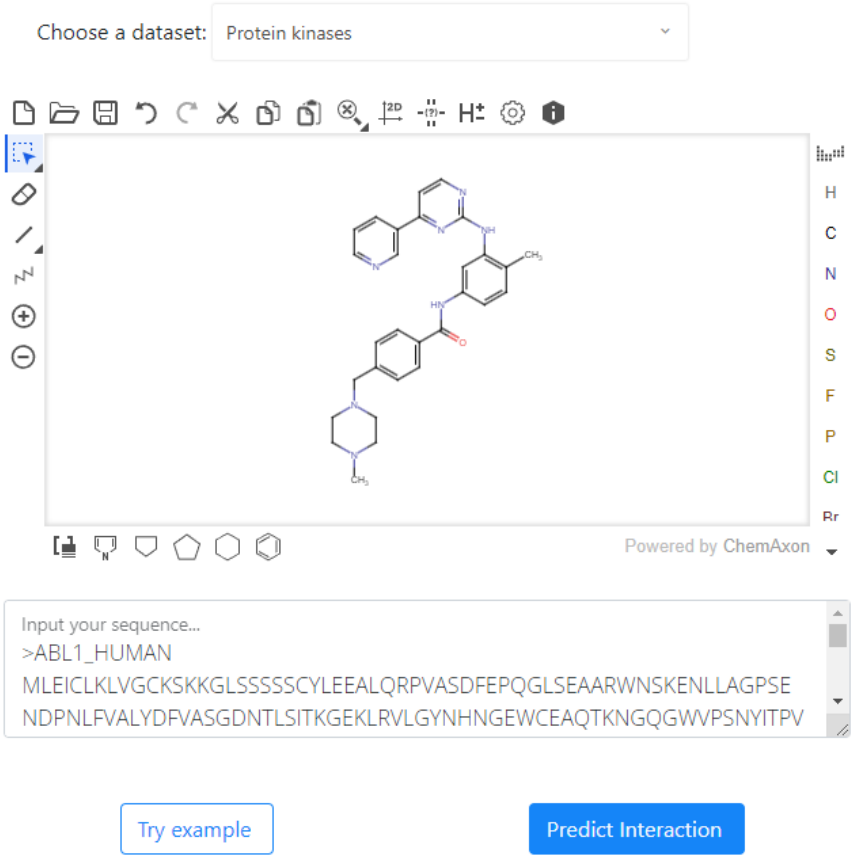

Рисунок 22. Интерфейс веб-сервиса с введенными тестовыми данными.

На выходе пользователь получает оценку *BE (Binding Estimation)*. Если *BE* выше нуля, то тестируемая пара рассматривается как связывающаяся и как не связывающаяся в противном случае.

Заключение

В настоящее время компьютерный прогноз белок-лигандных взаимодействий является важным этапом при поиске биологически активных соединений, в том числе при разработке новых лекарственных средств. Широкое применение нашли методы виртуального скрининга на основе сопоставления химических структур и анализа связи структура-активность (SAR). Существенная доля таких исследований направлена на выявление вероятных белков-мишеней, связывающих низкомолекулярные лиганды. В этом случае возможности прогностических методов существенно расширяются за счет включения характеристик мишеней в дополнение к структурным дескрипторам лигандов. Методы, построенные по этому принципу, объединяются термином «протеохемометрика».

В исследованиях по протеохемометрике используются различные способы описания белков – например, по оценкам сходства выровненных последовательностей; с помощью интегральных характеристик исследуемых белков, отражающих состав их последовательностей. Мы разработали программный комплекс, который включает в себя два модуля – PASS и SPrOS. SAR-метод PASS успешно применяется более двух десятилетий для прогноза биологической активности химических соединений [Filimonov и соавт., 2014], в том числе для выявления новых лекарственных мишеней [Pogodin, 2018]. Метод SPrOS основан на оценке локального сходства аминокислотных последовательностей [Karasev и соавт., 2016, Карасев и соавт., 2018, Karasev и соавт., 2020a, Karasev и соавт., 2020b, Karasev и соавт., 2020, Karasev и соавт., 2022]. Он был апробирован в данной работе как самостоятельный инструмент (позиционный режим, второй сценарий), так и в комбинации с PASS при реализации протеохемометрического похода (третий сценарий). Прогноз при совместной работе двух методов (третий сценарий) основан на идентификации локальных соответствий как между структурами лигандов, так и между аминокислотными последовательностями белков-мишеней.

В связи с включением дескрипторов аминокислотных последовательностей в прогностическую модель потребовались оценить область применимости подхода. С этой целью были собраны обучающие данные в соответствии с основными группами объектов, исследуемыми при компьютерной оценке белок-лигандных взаимодействий [Karasev и соавт., 2022]. Самая большая выборка содержала данные о взаимодействия лигандов с белками из разных структурно-функциональных групп без учета их гомологии. Четыре выборки содержали сведения об отдельных группах белков, представляющих классы перспективных лекарственных мишеней, а также об их лигандах. Третий тип данных представлял группы мутантных вариантов белков ВИЧ, различающиеся единичными аминокислотными заменами. Высокая точность прогноза была показана на выборках, которые представляли группы мишеней, отличающиеся по степени разнородности. Таким образом, мы продемонстрировали широкую область применимости, проведя испытания нашего подхода на различных данных, включающих как эталонные наборы, так и выборки, подготовленные на основе актуальных информационных ресурсов.

Разработанный нами программный комплекс, обладает широким функционалом и позволяет осуществлять прогноз на основе различных входных данных. Комплекс свободно доступен в виде веб-сервиса, что позволяет решать типовые задачи: определять новые вероятные мишени по химической структуре лиганда (SAR-моделирование), предсказывать новые лиганды для мишени по ее аминокислотной последовательности (биоинформатический анализ), оценивать взаимодействие белка и лиганда по дескрипторам обоих компонентов (протеохемометрика). Таким образом, наша технология открыта для широкого круга исследователей.

Выводы

1. Разработан оригинальный подход к извлечению данных из открытых источников о белок-лигандных взаимодействиях. Информация включает в себя структуры низкомолекулярных соединений, аминокислотные последовательности белков-мишеней, а также показатели взаимодействия для каждой пары. Полученные матрицы белок-лигандных взаимодействий включали от 400 и более значений для отдельных белковых семейств и до 2,7 миллионов значений в случае мишеней, неклассифицированных по филогении.
2. Создан метод для прогноза белок-лигандных взаимодействий на основе анализа локального сходства аминокислотных последовательностей и структур низкомолекулярных лигандов. Разработанный программный комплекс обеспечивает прогноз в соответствии с тремя типовыми сценариями: (1) новый лиганд – известный белок-мишень, (2) новый белок – известный лиганд, (3) новый белок – новый лиганд.
3. Проведена валидация метода на наборах, представляющих различные группы белков-мишеней, различающихся по филогенетическим отношениям. Высокая точность прогноза (IAP>0,85), достигнутая при всех трех сценариях, свидетельствует о широкой области применимости предложенного подхода, включая случаи с неполными обучающими данными.
4. Разработан свободно доступный в сети Интернет веб-сервис (http://www.way2drug.com/proteochemometrics/), позволяющий пользователю осуществлять прогноз во всех упомянутых сценариях для компьютерной оценки белок-лигандных взаимодействий.

Список работ, опубликованных по теме диссертации

Статьи в рецензируемых журналах

1. Karasev D. A., Sobolev B. N., Lagunin A. A., Filimonov D. A., Poroikov V. V. The method predicting interaction between protein targets and small-molecular ligands with the wide applicability domain // Computational biology and chemistry. – 2022. – V. 98, P. 107674.
2. Karasev D.A. Sobolev B.N., Lagunin A.A., Filimonov D.A., Poroikov V.V. Prediction of Protein-ligand Interaction Based on Sequence Similarity and Ligand Structural Features // International journal of molecular sciences. – 2020. – V. 21, № 21, P. 8152.
3. Karasev D.A. Sobolev B.N., Lagunin A.A., Filimonov D.A., Poroikov V.V. Prediction of Protein-Ligand Interaction Based on the Positional Similarity Scores Derived from Amino Acid Sequences // International journal of molecular sciences. – 2020. – V. 21(1), P. 24.
4. Карасев Д.А., Веселовский А.В., Лагунин А.А., Филимонов Д.А., Соболев Б.Н. Распознавание аминокислотных остатков, обуславливающих специфичное взаимодействие протеинкиназ с низкомолекулярными ингибиторами // Молекулярная биология – 2018. – Т. 52(3), С. 555–564
5. Karasev D. A., Veselovsky A. V., Oparina N. Y., Filimonov D. A., Sobolev, B. N. Prediction of amino acid positions specific for functional groups in a protein family based on local sequence similarity // Journal of molecular recognition. – 2016. – V. 29(4), P. 159–169.

Работы, опубликованные в сборниках материалов научных конференций

6. Karasev D.A., Sobolev B.N., Lagunin A.A., Filimonov D.A., Poroikov V.V. Predicting the protein-ligand interactions based on ligands' structures and proteins' sequences // The 13th International Conference on Bionifomatics of Genome Regulation and Structure/Systems Biology (BGRS\SB-2022). Novosibirsk, Russia, 2022. P. 345.
7. Карасев Д.А., Соболев Б.Н., Лагунин А.А., Филимонов Д.А., Поройков В.В. Прогноз связывания белков с низкомолекулярными лигандами на основе химических структур и аминокислотных последовательностей // VII Съезд биохимиков России и X Российский симпозиум "Белки и пептиды". Дагомыс, Россия, 2021. С. 173.
8. Карасев Д.А., Филимонов Д.А., Соболев Б.Н., Лагунин А.А. Прогноз белок-лигандных взаимодействий in silico // Международный форум: Биотехнология: состояние и перспективы развития. Москва, Россия, 2020. С. 254.

9. Karasev D.A., Filimonov D.A., B.N. Sobolev, Lagunin A.A. Detection of the protein targets for small molecular ligands with use of the local sequence similarity estimation // Moscow Conference on Computational Molecular Biology (MCCMB'19). Moscow, Russia, 2019.
10. Karasev D.A., Lagunin A.A., Filimonov D.A., Veselovsky A.V., Sobolev B.N. // 43rd FEBS Congress, Biochemistry Forever. Prague, Czech Republic, 2018. P. 287-288.
11. Karasev D.A., Savosina P.I., Veselovsky A.V., Filimonov D.A., Sobolev B.N. Identification of amino acid residues affecting on the specificity of interaction of protein kinases and small molecular inhibitors // Moscow Conference on Computational Molecular Biology (MCCMB'17). Moscow, Russia, 2017.
12. Карасев Д.А., Савосина П.И., Веселовский А.В., Филимонов Д.А., Соболев Б.Н. Определение лиганд-специфичных аминокислотных остатков в последовательностях протеинкиназ // VIII Российский симпозиум «белки и пептиды». Москва, Россия. 2017. С. 122.
13. Karasev D.A., Veselovsky A.V., Oparina N.Yu., Rudik A.V., Filimonov D.A., Sobolev B.N. Based on the local sequence similarity method for prediction of amino acid positions related to the protein-ligand specificity // The 10th International Conference on Bioinformatics of Genome Regulation and Structure/Systems Biology (BGRS\SB-2016). Novosibirsk, Russia, 2016. P. 117
14. Карасев Д.А., Веселовский А.В., Опарина Н.Ю., Рудик А.В., Филимонов Д.А., Соболев Б.Н. определение аминокислотных остатков, определяющих селективность ингибиторов к протеинкиназам человека // XXIII Российский национальный конгресс «Человек и лекарство». Москва, Россия 2016. С. 181.

Благодарности

- Автор выражает благодарность научному руководителю д.б.н., профессору РАН Алексею Александровичу Лагунину за поставленную актуальную интересную задачу и руководство на всех этапах исследования.
- Автор выражает благодарность к.б.н. Борису Николаевичу Соболеву за помощь при подготовке специализированной версии программы SPrOS и участие в интерпретации результатов.
- Автор выражает благодарность к.ф-м.н. Дмитрию Алексеевичу Филимонову за консультации при выработке общей концепции исследования и подготовку специализированной версии программы PASS.
- Автор выражает благодарность к.б.н. Дмитрию Сергеевичу Дружиловскому, к.б.н. Анастасии Владимировне Рудик и м.н.с. Никите Сергеевичу Ионову за помощь в разработке веб-сервиса.
- Автор выражает благодарность к.б.н. Опариной Нине Юрьевне за участие в создании набора искусственных последовательностей.
- Отдельную благодарность автор выражает своей семье, которые поддерживали его на протяжении выполнения диссертационной работы.
- Автор выражает благодарность Российскому Фонду Фундаментальных исследований за финансовую поддержку данной работы (гранты № 16-04-00491 и № 19-015-00374).

Список цитируемой литературы

1. Alexandrov K., Sobolev B., Filimonov D., Poroikov V. Recognition of protein function using the local similarity // Journal of bioinformatics and computational biology. 2008. V. 6. № 4. P. 709–725.
2. Alpaydin E. Introduction to machine learning // MIT Press, Cambridge. 2010.
3. Bender B.J., Gahbauer S., Luttens A., Lyu J., Webb C.M., Stein R.M., Fink E.A., Balius T.E., Carlsson J., Irwin J.J. Shoichet B.K. A practical guide to large-scale docking // Nat Protoc. 2021. V. 16. № 10. P. 4799-4832.
4. Blanco J.L., Porto-Pazos A.B., Pazos A., Fernandez-Lozano C. Prediction of high anti-angiogenic activity peptides in silico using a generalized linear model and feature selection // Sci Rep. 2018. V. 8. № 1. P. 1–11.
5. Bongers B.J., IJzerman A.P., Westen G.J.P. Proteochemometrics - recent developments in bioactivity and selectivity modeling // Drug Discov Today Technol. 2019 V. 32-33. P. 89-98.
6. Borrel A., Auerbach S.S., Houck K.A., Kleinstreuer N.C. Tox21 BodyMap: a webtool to map chemical effects on the human body // Nucleic Acids Res. 2020. V. 48 № W1. P. W472-W476.
7. Bradley D., Viéitez C., Rajeeve V., Selkrig J., Cutillas P.R., Beltrao P. Sequence and Structure-Based Analysis of Specificity Determinants in Eukaryotic Protein Kinases // Cell Rep. 2021. V. 12. № 34(2). P. 108602.
8. Chen H., Zhang Z. A semi-supervised method for drug-target interaction prediction with consistency in networks // PLoS ONE. 2013. V. 7. № 8(5). P. e62975.
9. Chen Z.H., You Z.H., Guo Z.H., Yi H.C., Luo G.X., Wang Y.B. Prediction of Drug-Target Interactions From Multi-Molecular Network Based on Deep Walk Embedding Model // Front Bioeng Biotechnol. 2020. V. 8. P. 338.
10. Chou K.C., Cai Y.D. Prediction of membrane protein types by incorporating amphipathic effects // J Chem Inf Model. 2005. V. 45. № 2 P. 407–413.
11. Christmann-Franck S., van Westen G.J., Papadatos G., Escudie F., Roberts A., Overington J.P., Domine D. Unprecedently Large-Scale Kinase Inhibitor Set Enabling the Accurate Prediction of Compound-Kinase Activities: A Way toward Selective Promiscuity by Design? // J Chem Inf Model. 2016. V. 56 № 9. P. 1654-1675.
12. Clark R., Fox P. Statistical variation in progressive scrambling // J Comput-Aided Mol Design. 2004. V. 18. P. 563–576.
13. Cohen P. Protein kinases--the major drug targets of the twenty-first century? // Nature reviews. Drug discovery. 2002. V. 1. № 4. P 309–315.
14. Cortes-Ciriano I., Murrell D.S., van Westen G.J., Bender A., Malliavin T.E. Prediction of the potency of mammalian cyclooxygenase inhibitors with ensemble proteochemometric modeling // J Cheminform. 2015. V. 7. P. 1-18.
15. Cortes-Ciriano I., Subramanian A.V. et al. Polypharmacology modelling using proteochemometrics (PCM): recent methodological developments, applications to target families, and future prospects // Medchemcomm. 2015 V. 6. P. 24–50
16. Cramer R.D., Wendt B. Pushing the boundaries of 3D-QSAR // J Comput Aided Mol Des. 2007 V. 21. P. 23-32.
17. Cronin M.T., Schultz T.W. Pitfalls in qsar // J Mol Struct (Thoechem) 2003. V. 622. P. 39–51.
18. Curtis C., Shah S.P., Chin S.-F., Turashvili G., Rueda O.M., Dunning M.J., Speed D., Lynch A.G., Samarajiwa S., Yuan Y. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups // Nature. 2012. V. 486. № 7403. P. 346–352.
19. Dana J.M., Gutmanas A., Tyagi N., Qi G., O'Donovan C., Martin M., Velankar S. SIFTS: updated Structure Integration with Function, Taxonomy and Sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins // Nucleic Acids Res. 2019. V. 8. № 47(D1). P. D482-D489.

20. Dara S., Dhamercherla S., Jadav S.S., Babu C.M., Ahsan M.J. Machine Learning in Drug Discovery: A Review // Artif Intell Rev. 2022. V. 55. № 3. P. 1947-1999.
21. Dardel F., Kepes F. Bioinformatics: Genomics and Post-Genomics // Wiley. 2006. p.54
22. Davis M.I., Hunt J.P., Herrgard S., Ciceri P., Wodicka L.M., Pallares G., Hocker M., Treiber D.K., Zarrinkar P.P. Comprehensive analysis of kinase inhibitor selectivity // Nat Biotechnol. 2011 V. 30. № 29(11). P. 1046-1051.
23. Dimitrov I., Garnev P., Flower D.R. et al. Peptide binding to the HLA-DRB1 supertype: a proteochemometrics analysis // Eur J Med Chem. 2010. V. 45. № 1. P. 236–243.
24. Dubchak I., Muchnik I., Holbrook S.R., Kim S.H. Prediction of protein folding class using global description of amino acid sequence // Proc Natl Acad Sci USA. 1995. V. 92. P. 8700-8704.
25. Duran-Frigola M., Mosca R., Aloy P. Structural systems pharmacology: the role of 3D structures in next-generation drug development // Chem Biol. 2013. V. 23. № 20(5). P. 674-84.
26. Feng Z.P., Zhang C.T. Prediction of membrane protein types based on the hydrophobic index of amino acids // J Protein Chem. 2000. V. 19. P. 262–275.
27. Filimonov D., Poroikov V. Chapter: Probabilistic approach in activity prediction // In book: Chemoinformatics Approaches to Virtual Screening. 2008 P.182-216.
28. Filimonov D., Poroikov V., Borodina Yu., Gloriozova T. Chemical similarity assessment through multilevel neighborhoods of atoms: definition and comparison with the other descriptors // Journal of Chemical Information and Computer Sciences. 1999. V. 39. № 4. P. 666-670.
29. Filimonov D.A. and Poroikov V.V. Chemoinformatics Approaches to Virtual Screening. // Cambridge. UK: Royal Society of Chemistry. 2008. P. 182-216.
30. Filimonov D.A., Lagunin A.A., Gloriozova T.A., Rudik A.V., Druzhilovskii D.S., Pogodin P.V., Poroikov V.V. Prediction of the biological activity spectra of organic compounds using the PASS online web resource // Chemistry of Heterocyclic Compounds. 2014. V. 50. № 3. P. 444-457.
31. 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. № 7. P. 1189-204.
32. Freyhult E., Prusis P., Lapinsh M., Wikberg J.E., Moulton V., Gustafsson M.G. Unbiased descriptor and parameter selection confirms the potential of proteochemometric modelling // BMC Bioinformatics. 2005. V. 10. № 6. P. 50.
33. Gao Q.B., Wang Z.Z., Yan C., Du Y.H. Prediction of protein subcellular location using a combined feature of sequence // FEBS Lett. 2005. V. 579. № 16. P. 3444–3448
34. García I., Munteanu C.R., Fall Y., Gómez G., Uriarte E., González-Díaz H. Qsar and complex network study of the chiral hmgr inhibitor structural diversity // Bioorganic Med Chem. 2009. V. 17. № 1. P. 165–175.
35. Geary R.C. The Contiguity Ratio and Statistical Mapping // The Incorporated Statistician. 1954. V. 5 № 3. P. 115–145.
36. Gedeck P., Rohde B., Bartels C. QSAR – how good is it in practice? Comparison of descriptor sets on an unbiased cross section of corporate data sets // J Chem Inf Model. 2006. V. 46. № 5. P. 1924-1936.
37. Geppert H., Humrich J., Stumpfe D., Gärtner T., Bajorath J. Ligand prediction from protein sequence and small molecule information using support vector machines and fingerprint descriptors // J Chem Inf Model. 2009. V. № 49. P. 767-779.
38. Ghosh A.K., Gemma S. Carbonic anhydrase inhibitors for the treatment of glaucoma: Design and discovery of dorzolamide, in structurebased design of drugs and other bioactive molecules // Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA, 2014, C. 19.
39. Giblin K.A., Hughes S.J., Boyd H., Hansson P., Bender A. Prospectively validated proteochemometric models for the prediction of small-molecule binding to bromodomain proteins // J Chem Inf Model. 2018. V. 58. P. 1870-1888.
40. Goodarzi M., Dejaegher B., Vander Heyden Y. Feature selection methods in QSAR studies // J AOAC Int. 2012. V. 95 № 3. P. 636-651.

41. Gozalbes R., Doucet J. P., Derouin F. Application of topological descriptors in QSAR and drug design: history and new trends // Current drug targets. Infectious disorders. 2002. V. 2. № 1. P. 93–102.
42. Guha R., Van Drie J.H. Structure--activity landscape index: identifying and quantifying activity cliffs // J Chem Inf Model. 2008. V. 48. № 3. P. 646-658
43. Hanash S. HUPO initiatives relevant to clinical proteomics // Mol Cell Proteomics. 2004. V. 3. № 4. P. 298-301.
44. Hansch C., Maloney P.P., Fujita T., Muir R.M. Correlation of biological activity of phenoxyacetic acids with hammett substituent constants and partition coefficient // Nature. 1962. P. 194. № 4824. P. 178–180.
45. Hariri S., Ghasemi J.B., Shirini F., Rasti B. Probing the origin of dihydrofolate reductase inhibition via proteochemometric modeling // J Chemom. 2019. V. 33. P. e3090.
46. Hariri S., Rasti B., Mirpour M., Vaghar-Lahijani G., Attar F., Shiri F. Structural insights into the origin of phosphoinositide 3-kinase inhibition // Struct Chem. 2020. V. 31. P. 1-18.
47. Hartman G.D., Egbertson M.S., Halczenko W., Laswell W.L., Duggan M.E., Smith R.L., Naylor A.M., Manno P.D., Lynch R.J., Zhang G. et al. Non-peptide fibrinogen receptor antagonists. 1. Discovery and design of exosite inhibitors // J Med Chem. 1992. V. 35 P. 4640–4642.
48. Horne D.S. Prediction of protein helix content from an autocorrelation analysis of sequence hydrophobicities // Biopolymers. 1988. V. 27. P. 451–477.
49. Huang Q., Jin H.X., Liu Q. et al. Proteochemometric modeling of the bioactivity spectra of HIV-1 protease inhibitors by introducing protein-ligand interaction fingerprint // PloS One 2012. V. 7. № 7. P. e41698.
50. Huang Q., Jin H.X., Liu Q. et al. Proteochemometric modeling of the bioactivity spectra of HIV-1 protease inhibitors by introducing protein-ligand interaction fingerprint // PloS One 2012. V. 7. № 7. P. e4169808.
51. Junaid M., Lapins M., Eklund M et al. Proteochemometric modeling of the susceptibility of mutated variants of the HIV-1 virus to reverse transcriptase inhibitors // PloS One. 2010 V. 5. № 12.
P. e14353.
52. Junaid M., Lapins M., Eklund M., Spjuth O., Wikberg J.E. Proteochemometric modeling of the susceptibility of mutated variants of the HIV-1 virus to reverse transcriptase inhibitors // PLoS One. 2010. V. 5. P. e14353.
53. Kalliokoski, T., Kramer, C., Vulpetti, A. Quality Issues with Public Domain Chemogenomics Data // Molecular informatics. 2013. V. 32. P. 898–905.
54. Karasev D. A., Sobolev B. N., Lagunin A. A., Filimonov D. A., Poroikov V. V. The method predicting interaction between protein targets and small-molecular ligands with the wide applicability domain // Computational biology and chemistry. 2022. V. 98. P. 107674.
55. Karasev D. A., Veselovsky A. V., Oparina N. Y., Filimonov D. A., Sobolev, B. N. Prediction of amino acid positions specific for functional groups in a protein family based on local sequence similarity // Journal of molecular recognition. 2016. V. 29. № 4, P. 159–169.
56. Karasev D.A., Sobolev B.N., Lagunin A.A., Filimonov D.A., Poroikov V.V. Prediction of Protein-ligand Interaction Based on Sequence Similarity and Ligand Structural Features // International journal of molecular sciences. 2020. V. 21. № 21. P. 8152.
57. Karasev D.A., Sobolev B.N., Lagunin A.A., Filimonov D.A., Poroikov V.V. Prediction of Protein-Ligand Interaction Based on the Positional Similarity Scores Derived from Amino Acid Sequences // International journal of molecular sciences. 2020. V. 21. № 1. P. 24.
58. Kim P., Winter R., Clevert D.A. Deep protein-ligand binding prediction using unsupervised learned representations // ChemRxiv. 2020.
59. Kontijevskis A., Komorowski J., Wikberg J.E. Generalized proteochemometric model of multiple cytochrome p450 enzymes and their inhibitors // J Chem Inf Model. 2008 V. 48. № 9. P. 1840-50.
60. Lapins M., Eklund M., Spjuth O. et al. Proteochemometric modeling of HIV protease susceptibility // BMC Bioinformatics. 2008. V. 9. P. 181.

61. Lapins M., Wikberg J.E. Kinome-wide interaction modelling using alignment-based and alignment-independent approaches for kinase description and linear and non-linear data analysis techniques // BMC Bioinformatics. 2010. V. 11. P. 339
62. Lapins M., Wikberg J.E. Kinome-wide interaction modelling using alignment-based and alignment-independent approaches for kinase description and linear and non-linear data analysis techniques // BMC bioinformatics. 2010. V. 11. P. 339.
63. Lapins M., Wikberg J.E.S. Kinome-wide interaction modelling using alignment-based and alignment-independent approaches for kinase description and linear and non-linear data analysis techniques // BMC Bioinformatics. 2010. V. 11. P. 339.
64. Lapins M., Wikberg J.E.S. Proteochemometric modeling of drug resistance over the mutational space for multiple HIV protease variants and multiple protease inhibitors // J Chem Inf Model 2009. V. 49. P. 1202–10.
65. Lapinsh M., Prusis P., Gutcaits A., Lundstedt T., Wikberg J.E. Development of proteo-chemometrics: a novel technology for the analysis of drug-receptor interactions // Biochim Biophys Acta. 2001. V. 1525. P. 180-190.
66. Lapinsh M., Prusis P., Lundstedt T. et al. Proteochemometrics modeling of the interaction of amine G-protein coupled receptors with a diverse set of ligands // Mol Pharmacol. 2002. V. 61. P. 1465–1475.
67. Lapinsh M., Prusis P., Mutule I., Mutulis F., Wikberg, J.E. QSAR and proteo-chemometric analysis of the interaction of a series of organic compounds with melanocortin receptor subtypes // Journal of medicinal chemistry. 2003. V. 46. № 13. P. 2572–2579.
68. Lapinsh M., Prusis P., Uhlen S et al. Improved approach for proteochemometrics modeling: application to organic compound - amine G protein-coupled receptor interactions // Bioinformatics. 2005. V. 21. № 23. P. 4289–4296
69. Lapinsh M., Prusis P., Uhlén S., Wikberg J.E.S. Improved approach for proteochemometrics modeling: application to organic compound—amine G protein-coupled receptor interactions // Bioinformatics. 2005. V. 21. № 23. P. 4289–4296.
70. Lapinsh M., Veiksina S., Uhlén S., Petrovska R., Mutule I., Mutulis F., Yahorava S., Prusis P., Wikberg J.E. Proteochemometric mapping of the interaction of organic compounds with melanocortin receptor subtypes // Mol Pharmacol. 2005. V. 67. № 1. P. 50-59.
71. Lenselink E.B., ten Dijke N., Bongers B., Papadatos G., van Vlijmen H.W.T., Kowalczyk W. et al. Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set // J Cheminform. 2017. V. 9. P. 45.
72. Li Z.R., Lin H.H., Han L.Y., Jiang L., Chen X., Chen Y.Z. PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence // Nucleic Acids Res. 2006. V. 34. P. W32-W37.
73. Lipinski C.A., Lombardo F., Dominy B.W., Feeney P.J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings // Advanced Drug Delivery Reviews. 2001. V. 46 P. 3–26.
74. Lipman D.J., Pearson W.R. Rapid and sensitive protein similarity searches // Science. 1985. V. 227. № 4693. P. 1435-1441.
75. Liu Y., Tang S., Fernandez-Lozano C., Munteanu C.R., Pazos A., Yu Y.Z., Tan Z., González-Díaz H. Experimental study and random forest prediction model of microbiome cell surface hydrophobicity // Expert Syst Appl. 2017. V. 72. P. 306–316.
76. Liu Y., Wu M., Miao C., Zhao P., Li X.L. Neighborhood regularized logistic matrix factorization for drug-target interaction prediction // PLoS Comput. Biol. 2016. V. 12. № 2. P. e1004760.
77. Liu Z., Li Y., Han L., Li J., Liu J., Zhao Z., Nie W., Liu Y., Wang R. PDB-wide collection of binding data: current status of the PDBbind database // Bioinformatics. 2015 V. 31. № 3. P. 405-412.
78. Manoharan P., Chennoju K., Ghoshal N. Target specific proteochemometric model development for BACE1 - protein flexibility and structural water are critical in virtual screening // Mol Biosyst. 2015 V. 11. № 7. P. 1955-1972.

79. Mauri A., Consonni V., Pavan M., Todeschini R. Dragon software: An easy approach to molecular descriptor calculations // MATCH Communications in Mathematical and in Computer Chemistry. 2006. V. 56. P. 237–248.
80. Mazanetz M.P., Marmon R.J., Reisser C.B., Morao I. Drug discovery applications for KNIME: an open source data mining platform // Curr Top Med Chem. 2012 V. 12. № 18. P. 1965-1979.
81. Medina-Franco J.L., Martinez-Mayorga K., Bender A., Marin R. M., Giulianotti M. A., Pinilla C., Houghten R. A. Characterization of activity landscapes using 2D and 3D similarity methods: consensus activity cliffs. // J. Chem. Inf. Model. V. 2009. V. 49. P. 477–491.
82. Mendez D., Gaulton A., Bento A.P., Chambers J., De Veij M., Félix E., Magariños M.P., Mosquera J.F., Mutowo P., Nowotka M., Gordillo-Marañón M., Hunter F., Junco L., Mugumbate G., Rodriguez-Lopez M., Atkinson F., Bosc N., Radoux C.J., Segura-Cabrera A., Leach A.R. ChEMBL: towards direct deposition of bioassay data // Nucleic Acids Res. 2019/ V. 8. № 47(D1). P. D930-D940.
83. Mistry J., Chuguransky S., Williams L., Qureshi M., Salazar G.A., Sonnhammer E.L.L., Tosatto S.C.E., Paladin L., Raj S., Richardson L.J., Finn R.D., Bateman A. Pfam: The protein families database in 2021 // Nucleic Acids Res. 2021 V. 8. № 49(D1). P. D412-D419.
84. Munteanu C.R., Fernández-Blanco E., Seoane J.A., Izquierdo-Novo P., Angel Rodriguez-Fernandez J., Maria Prieto-Gonzalez J., Rabunal J.R., Pazos A. Drug discovery and design for complex diseases through qsar computational methods // Current Pharmaceutical Des. 2010. V. 16. № 24. P. 2640–2655.
85. 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., Curtarolo S., Fourches D., Cohen Y., Aspuru-Guzik A., Winkler D.A., Agrafiotis D., Cherkasov A., Tropsha A. QSAR without borders // Chem Soc Rev. 2020. V. 7. № 49. P. 3525-3564.
86. Murgueitio M.S., Bermudez M., Mortier J., Wolber G.. In silico virtual screening approaches for anti-viral drug discovery // Drug Discov. Today. 2012. V. 9. P. 219–225.
87. Murrell D.S., Cortes-Ciriano I., van Westen G.J.P., Stott I.P., Bender A., Malliavin T.E., Glen R.C. Chemically Aware Model Builder (camb): an R package for property and bioactivity modelling of small molecules // J Cheminform. 2015. V. 7. № 45.
88. Nabu S., Nantasenamat C., Owasirikul W., Lawung R., Isarankura-Na-Ayudhya C., Lapins M., Prachayasittikul V. Proteochemometric model for predicting the inhibition of penicillin-binding proteins // Journal of Computer-Aided Molecular Design. 2014. V. 29. № 2. P. 127–141.
89. Nazarshodeh E., Sheikhpour R., Gharaghani S., Sarram M.A. A novel proteochemometrics model for predicting the inhibition of nine carbonic anhydrase isoforms based on supervised Laplacian score and k-nearest neighbour regression // SAR QSAR Environ Res. 2018. V. 29. P. 419-437.
90. Neves B.J., Braga R.C., Melo-Filho C.C., Moreira-Filho J.T., Muratov E.N., Andrade C.H. QSAR-Based Virtual Screening: Advances and Applications in Drug Discovery // Front Pharmacol. 2018. V. 9. P. 1275.
91. Ning X., Rangwala H., Karypis G. Multi-assay-based structure-activity relationship models: improving structure-activity relationship models by incorporating activity information from related targets // J Chem Inf Model. 2009. V. 49. № 11. P. 2444-2456.
92. Oprea T.I., Mestres J. Drug repurposing: far beyond new targets for old drugs // AAPS J. 2012. V. 14 № 4. P. 759-63.
93. Ozturk H., A. Ozgur A., Ozkirimli E. DeepDTA: deep drug-target binding affinity prediction Bioinformatics. 2018. V. 34. P. i821-i829.
94. Pahikkala T., Airol, A., Pietilä S., Shakyawar, S., Szwajda A., Tang J., Aittokallio, T. Toward more realistic drug-target interaction predictions // Brief Bioinform. 2015. V. 16. № 2. P. 325-37.
95. Pastor M., Cruciani G., McLay I., Pickett S., Clementi S. GRid-INdependent descriptors (GRIND): a novel class of alignment-independent three-dimensional molecular descriptors // J Med Chem. 2000. V. 43. № 17. P. 3233-3243.

96. Patel H., Ihlenfeldt W.D., Judson P.N., Moroz Y.S., Pevzner Y., Peach M.L, Delannée V., Tarasova N.I., Nicklaus M.C. SAVI, in silico generation of billions of easily synthesizable compounds through expert-system type rules // Sci Data. 2020. V. 7. № 1. P. 384.
97. Paysan-Lafosse T., Blum M., Chuguransky S., Grego T., Pinto B.L., Salazar G.A., Bileschi M.L., Bork P., Bridge A., Colwell L., Gough J., Haft D.H., Letunić I., Marchler-Bauer A., Mi H., Natale D.A., Orengo C.A., Pandurangan A.P., Rivoire C., Sigrist C.J.A., Sillitoe I., Thanki N., Thomas P.D., Tosatto S.C.E., Wu C.H., Bateman A. InterPro in 2022 // Nucleic Acids Res. 2023. V. 51. № D1. P. D418-D427.
98. Peter W. Rose, Chunxiao Bi, Wolfgang F. Bluhm, Cole H. Christie, Dimitris Dimitropoulos, Shuchismita Dutta, Rachel K. Green, David S. Goodsell, Andreas Prlić, Martha Quesada, Gregory B. Quinn, Alexander G. Ramos, John D. Westbrook, Jasmine Young, Christine Zardecki, Helen M. Berman, Philip E. Bourne, The RCSB Protein Data Bank: new resources for research and education, Nucleic Acids Research, Volume 41, Issue D1, 1 January 2013, Pages D475–D482.
99. Pogodin P.V., Lagunin A.A., Filimonov D.A., Nicklaus M.C., Poroikov V.V. Improving (Q)SAR predictions by examining bias in the selection of compounds for experimental testing // SAR QSAR Environ Res. 2019. V. 30. № 10. P. 759-773.
100. Prusis P., Junaid M., Petrovska R. et al. Design and evaluation of substrate-based octapeptide and non substrate-based tetrapeptide inhibitors of dengue virus NS2B-NS3 proteases // Biochem Biophys Res Commun. 2013. V. 434. № 4. P. 767–772.
101. Prusis P., Lapins M., Yahorava S. et al. Proteochemometrics analysis of substrate interactions with dengue virus NS3 proteases // Bioorgan Med Chem. 2008. V.16. № 20. P. 9369–9377.
102. Prusis P., Muceniece R., Andersson P., Post C., Lundstedt T., Wikberg J.E. PLS modeling of chimeric MS04/MSH-peptide and MC1/MC3-receptor interactions reveals a novel method for the analysis of ligand-receptor interactions // Biochim Biophys Acta. 2001. V. 12. P. 350-357.
103. Pundir S., Martin M.J., O'Donovan C. UniProt Protein Knowledgebase // Methods Mol Biol. 2017 V. 1558. P.41-55.
104. Qiu T., Qiu J., Feng J., Wu D., Yang Y., Tang K., Cao Z., Zhu R. The recent progress in proteochemometric modelling: focusing on target descriptors, cross-term descriptors and application scope // Briefings in Bioinformatics. V. 18. № 1. P. 125–136.
105. Qiu T., Wu D., Qiu J., Cao Z. Finding the molecular scaffold of nuclear receptor inhibitors through high-throughput screening based on proteochemometric modelling // J Cheminform. 2018. V. 10. P. 21.
106. Raevsky O.A. Physicochemical descriptors in property-based drug design // Mini Rev Med Chem. 2004. V. 4. № 10. P. 1041-1052.
107. Rasti B., Schaduangrat N., Shahangian S.S., Nantasenamat C. Exploring the origin of phosphodiesterase inhibition: via proteochemometric modeling // RSC Adv. 2017. V. 7. P. 28056-28068.
108. Rasti B., Schaduangrat N., Shahangian S.S., Nantasenamat C. Exploring the origin of phosphodiesterase inhibition: via proteochemometric modeling // RSC Adv. 2017. V. 7 P. 28056-28068.
109. Rasti B., Shahangian S.S. Proteochemometric modeling of the origin of thymidylate synthase inhibition // Chem Biol Drug Des. 2018. V. 91 P. 1007-1016.
110. Reker D., Schneider P., Schneider G., Brown J.B. Active learning for computational chemogenomics. // Future Med Chem. 2017. V. 9. № 4. P 381-402.
111. Rhee S.Y., Gonzales M.J., Kantor R., Betts B.J., Ravela J., Shafer R.W. Human immunodeficiency virus reverse transcriptase and protease sequence database // Nucleic Acids Res. 2003. V. 1. № 31(1). P. 298-303.
112. Richard A.M., Huang R., Waidyanatha S., Shinn P., Collins B.J., Thillainadarajah I., Grulke C.M., Williams A.J., Lougee R.R., Judson R.S., Houck K.A., Shobair M., Yang C., Rathman J.F., Yasgar A., Fitzpatrick S.C., Simeonov A., Thomas R.S., Crofton K.M., Paules R.S., Bucher J.R.,

Austin C.P., Kavlock R.J., Tice R.R. The Tox21 10K Compound Library: Collaborative Chemistry Advancing Toxicology // Chem Res Toxicol. 2021 V. 15. № 34. P. 189-216.
113. Riera-Fernández P., Munteanu C.R., Dorado J., Martin-Romalde R., Duardo-Sanchez A., Gonzalez-Diaz H. From chemical graphs in computer-aided drug design to general markov-galvez indices of drug-target, proteome, drug-parasitic disease, technological, and social-legal networks // Current Computer-aided Drug Des. 2011. V. 7. № 4. P. 315–337.
114. Rogers D., Hahn J.M. Extended-Connectivity Fingerprints // Chem. Inf. Model. 2010. V. 50 № 5. P. 742–754.
115. Romond E.H., Perez E.A., Bryant J., Suman V.J., Geyer C.E., Jr, Davidson N.E., Tan-Chiu E., Martino S., Paik S., Kaufman P.A. Trastuzumab plus adjuvant chemotherapy for operable her2-positive breast cancer // N Engl J Med. 2005 V. 353. № 16. P. 1673–1684.
116. Rose P.W., Bi C., Bluhm W.F., Christie C.H., Dimitropoulos D., Dutta S., Green R.K., Goodsell D.S., Prlic A., Quesada M., Quinn G.B., Ramos A.G., Westbrook J.D., Young J., Zardecki C., Berman H.M., Bourne P.E. The RCSB Protein Data Bank: new resources for research and education // Nucleic Acids Res. 2013 V. 41 (Database issue). P. D475-82
117. Sandberg M., Eriksson L., Jonsson J., Sjostrom M., Wold S. New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids // Journal of medicinal chemistry. 1998. V. 41. P. 2481–2491.
118. Santos R., Ursu O., Gaulton A., Bento A.P., Donadi R.S., Bologa C.G., Karlsson A., Al-Lazikani B., Hersey A., Oprea T.I., Overington J.P. A comprehensive map of molecular drug targets // Nat Rev Drug Discov. 2017. V. 16 № 1. P. 19-34.
119. Savosina P.I., Druzhilovskii D.S., Poroikov V.V. COVID-19: Analysis of Drug Repositioning Practice // Pharm Chem J. 2021. V. 54. № 10. P. 989-996.
120. Shafer R.W. Rationale and Uses of a Public HIV Drug-Resistance Database // Journal of Infectious Diseases. 2006 V.194.
121. Shaikh N., Sharma M., Garg P. An improved approach for predicting drug–target interaction: proteochemometrics to molecular docking // Molecular BioSystems. 2016. V. 3.
122. Shar P.A., Tao W., Gao S., Huang C., Li B., Zhang W., et al. Pred-binding: largescale protein–ligand binding affinity prediction // J Enzyme Inhib Med Chem. 2016. V. 31. P. 1443–1450.
123. Shi J.Y., Zhang A.Q., Zhang S.W., Mao K.T., Yiu S.M. A unified solution for different scenarios of predicting drug-target interactions via triple matrix factorization // BMC Syst. Biol. 2018. V. 12. № 136.
124. Shi J.Y., Zhang A.Q., Zhang S.W., Mao K.T., Yiu S.M. A unified solution for different scenarios of predicting drug-target interactions via triple matrix factorization // BMC Syst. Biol. 2018. V. 12. P. 136.
125. Shirvani P., Fassihi A. Molecular modelling study on pyrrolo [2, 3-b] pyridine derivatives as c-met kinase inhibitors, a combined approach using molecular docking, 3D-qsar modelling and molecular dynamics simulation // Mol Simul. 2020. P. 1265-1280.
126. Simeon S., Spjuth O., Lapins M., Nabu S., Anuwongcharoen N., Prachayasittikul V. et al. Origin of aromatase inhibitory activity via proteochemometric modeling // PeerJ. 2016. V. 4. P. e1979.
127. Sliwoski G., Kothiwale S., Meiler J., Lowe E.W. Jr. Computational methods in drug discovery // Pharmacol Rev. 2013. V. 66. № 1. P. 334-395.
128. Sorgenfrei F.A., Fulle S., Merget B.. Kinome-Wide Profiling Prediction of Small Molecules // ChemMedChem. 2018. V. 13. № 6. P. 495-499.
129. Sriram K., Insel P.A. G Protein-Coupled Receptors as Targets for Approved Drugs: How Many Targets and How Many Drugs? // Mol Pharmacol. 2018. V. 93 № 4. P. 251-258.
130. Stroembergsson H., Daniluk P., Kryshtafovych A. et al. Interaction model based on local protein substructures generalizes to the entire structural enzyme-ligand space // J Chem Inf Model. 2008. V. 48. № 11. P. 2278–2288.

131. Strombergsson H., Kryshtafovych A., Prusis P. et al. Generalized modeling of enzyme-ligand interactions using proteochemometrics and local protein substructures // Proteins. 2006. V. 65. № 3. P. 568–579.
132. Strombergsson H., Lapins M., Kleywegt G.J. et al. Towards proteome-wide interaction models using the proteochemometrics approach // Mol Inform. 2010. V. 29 V. 499–508.
133. Suay-Garcia B., Bueso-Bordils J.I., Falcó A., Pérez-Gracia M.T., Antón-Fos G., Alemán-López P., Quantitative structure–activity relationship methods in the discovery and development of antibacterials // Wiley Interdisciplinary Reviews: Computational Molecular Science. 2020 P. e1472.
134. Sun D., Gao W., Hu H., Zhou S. Why 90% of clinical drug development fails and how to improve it? // Acta pharmaceutica Sinica. 2022. V. 12. № 7, P. 3049–3062.
135. Sviatopolk-Mirsky F.P., de Cássia Ruy P., Oliveira G., Coimbra R.S. Assessing the efficiency of multiple sequence alignment programs // Algorithms for Molecular Biology 2014. V. 9.P. 4.
136. Talele T.T., Khedkar S.A., Rigby A.C. Successful applications of computer aided drug discovery: moving drugs from concept to the clinic // Curr Top Med Chem. 2010. V. 10. P. 127–141.
137. Tetko I.V., Maran U., Tropsha A. Public (Q)SAR Services, Integrated Modeling Environments, and Model Repositories on the Web: State of the Art and Perspectives for Future Development // Mol Inform. 2017. V. 36. № 3.
138. Tresadern G., Trabanco A.A., Pérez-Benito L., Overington J.P., Van Vlijmen H.W.T., van Westen G.J.P. Identification of allosteric modulators of metabotropic glutamate 7 receptor using proteochemometric modeling // J Chem Inf Model. 2017. V. 57. P. 2976-2985.
139. Vamathevan J., Clark D., Czodrowski P., Dunham I., Ferran E., Lee G., Li B., Madabhushi A., Shah P., Spitzer M., Zhao S. Applications of machine learning in drug discovery and development // Nat Rev Drug Discov. 2019. V. 18. P. 463-477.
140. Van Drie J.H. Computer-aided drug design: the next 20 years // J Comput Aided Mol Des. 2007. V. 21 P. 591–601.
141. Van Laarhoven T., Marchiori, E. Predicting Drug-Target Interactions for New Drug Compounds Using a Weighted Nearest Neighbor Profile // PLoS ONE 2013. V. 8. № 6. P. e66952.
142. Velankar S., Burley S.K., Kurisu G., Hoch J.C., Markley J.L. The Protein Data Bank Archive // Methods Mol Biol. 2021. V. 2305. P. 3-21.
143. Vijayakrishnan R. Structure-based drug design and modern medicine // J Postgrad Med. 2009. V. 55. P. 301–304.
144. Wade R.C., Salo-Ahen O.M.H. Molecular Modeling in Drug Design // Molecules. 2019. V. 24. № 2 P. 321.
145. Wang H., Zheng H. Model Validation, Machine Learning // Encyclopedia of Systems Biology. 2013. P. 1406–1407.
146. Wawer M., Peltason L., Bajorath J. Elucidation of structure-activity relationship pathways in biological screening data // J Med Chem. 2009. V. 26. № 52. P. 1075-1080.
147. Westen G. J. P., Wegner J.K, IJzerman A.P., Vlijmenab H.W.T., Bender A. Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets // Med. Chem. Commun. 2011. V. 2. P. 16-30.
148. Wold S., Jonsson J., Sjostrom M., Sandberg M., Rannar S. DNA and peptide sequences and chemical processes multivariately modelled by principal component analysis and partial least-squares projections to latent structures // Anal Chim Acta. 1993. V. 277. P. 239-252.
149. Xia Z., Wu L.Y., Zhou X.,Wong S.T. Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces // BMC Syst. Biol. 2010. V. 4.
150. Yap C.W. PaDEL-Descriptor: An open source software to calculate molecular descriptors and fingerprints. // Journal of Computational Chemistry. 2011. V. 32. № 7 P. 1466-1474.
151. Yogesh R. Python: Simple though an Important Programming language // IRJET. 2019. V. 6. № 2. P. 1856—1858.

152. Zakharov A.V., Zhao T., Nguyen D.T., Peryea T., Sheils T., Yasgar A. et al. Novel consensus architecture to improve performance of large-scale multitask deep learning QSAR models // J Chem Inf Model. 2019. V. 59 P. 4613-4624.
153. Zheng X., Ding H., Mamitsuka H., Zhu S. Collaborative matrix factorization with multiple similarities for predicting drug-target interactions // Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, IL, 2013.
154. Zhu Y., Hu X. Molecular Recognition of FDA-Approved Small Molecule Protein Kinase Drugs in Protein Kinases // Molecules (Basel, Switzerland). 2022. V. 27. № 20. P. 7124.
155. Карасев Д.А., Веселовский А.В., Лагунин А.А., Филимонов Д.А., Соболев Б.Н. Распознавание аминокислотных остатков, обуславливающих специфичное взаимодействие протеинкиназ с низкомолекулярными ингибиторами // Молекулярная биология. 2018. Т. 52. № 3, С. 555–564