И.А. Моргунов, И.Д. Никитин, В.Т. Ткаченко, М.В. Федоров

ХЕМОИНФОРМАТИКА И ВЫЧИСЛИТЕЛЬНАЯ ТОКСИКОЛОГИЯ

Посвящаем Светлой Памяти
Игоря Анатольевича Кириллова
и Ильи Владимировича Поликарпова.

МОСКВА
2025

УДК 54; 615
ББК 24в615.2
М 79

Рецензент:
д.б.н., проф. А.В. Поздеев

И.А. Моргунов, И.Д. Никитин, В.Т. Ткаченко, М.В. Федоров / **Хемоинформатика и вычислительная токсикология** – М.: РАН, 2025. – с. 128

Монография предназначена для студентов, аспирантов, преподавателей и научных сотрудников в области химических наук и химической безопасности. В частности, медицинской химии; аналитической химии; химической технологии; хемоинформатики; инфохимии; токсикологии, экологии и природопользования; а также фармакологии и смежных научных областей. Эту книгу можно рассматривать как учебник, который станет надежным проводником в бурно развивающуюся область цифровой химии, которая за последние десятилетия сильно поменяла подход к классическим исследованиям в области прогнозирования свойств химических веществ, оценке их токсикологических рисков и пригодности как лекарственных кандидатов. В данном пособии описаны современные подходы по анализу различной химической информации, а также методы, позволяющие эффективно искать, обрабатывать и интерпретировать большие объемы данных, которые используются в хемоинформатике. Помимо прочего, монография охватывает ключевые концепции молекулярного моделирования и машинного обучения. Эта книга подойдет для быстрого старта и погружения в основные аспекты хемоинформатики, давая понимание широкому кругу читателей о том, как была устроена раньше и как развивается сейчас одна из самых многообещающих наук нашего столетия, постепенно вытесняющая классические методы исследований даже со сцены вручения Нобелевских премий.

Цель книги – не заменить собой глубокие специализированные обзоры и монографии (ссылки на которые приведены по ходу изложения и в списке литературы), а дать читателю прочный концептуальный фундамент и системное понимание:

1. Объяснить симбиоз между областями: показать, какие инструменты и данные химической информатики необходимы и используются для построения и применения моделей вычислительной токсикологии.

2. Структурировать знание: представить ключевые концепции, методологии и задачи вычислительной токсикологии (прогнозирование параметров, приоритизация, механизмы) в логической последовательности, связанной с основами химической информатики.

3. Обозначить контекст: кратко осветить движущие силы развития области (этика, экономика, регуляторные требования, объем "химического пространства").

4. Указать путь для углубления знаний: каждая тема, затронутая в книге (будь то конкретные типы дескрипторов, алгоритмы машинного обучения или модели конкретных видов токсичности), представляет собой огромное поле для самостоятельного изучения.

Кроме того, данное пособие будет полезно и состоявшимся специалистам в этой области, поскольку содержит систематизацию большого количества информации, актуальной в исследованиях подобного характера. Таким образом, необходимость в этой вводной монографии была продиктована самой природой области: ее колоссальным объемом, быстрым развитием и сложной интердисциплинарностью. Она написана с целью помочь читателю осуществлять эффективную навигацию в море существующих публикаций и осознанно выбирать направления для глубокого погружения в конкретные аспекты химической информатики и вычислительной токсикологии, столь важные для безопасности нашего будущего.

Эта книга была создана коллективными усилиями группы специалистов в различных областях химии, молекулярного моделирования, медицины и машинного обучения. От чего и получилась разнообразной и дополняющей каждое из этих направлений до (надеемся) единой цельной концепции, движущей хемоинформатику к новым вершинам. Приятного прочтения!

ISBN 978-5-907645-73-8

СОДЕРЖАНИЕ

1. ВВЕДЕНИЕ

Уважаемый читатель! Если эта книга оказалась у Вас в руках, значит, прямо сейчас Вы ступаете на порог одной из самых динамично развивающихся наук нашего времени – **химической информатики или хемоинформатики**.

Перед Вами открывается мир, где мощь компьютерных алгоритмов встречается с бесконечным разнообразием молекулярных структур. Мир, который родился на острейшем стыке дисциплин: медицинской химии, постоянно ищущей пути к новым лекарствам; математической статистики, выявляющей скрытые закономерности; вычислительной токсикологии, предупреждающей о рисках; физической химии, описывающей взаимодействия на атомарном уровне, а также многих других.

Эта область открывает новые горизонты в понимании и предсказании свойств химических соединений, позволяя решать сложнейшие задачи современного научного и прикладного характера. Химическая информатика – это не просто набор инструментов и методов, это целая философия работы с данными, которая помогает исследователям создавать эффективные модели, ускорять разработку лекарств, прогнозировать токсичность веществ и находить новые решения в самых разных областях молекулярных наук.

Как найти ту самую, единственную молекулу-ключ среди миллионов и миллиардов возможных? Как предсказать, будет ли она лечить или, наоборот, навредит? Как ускорить долгий и дорогой путь от идеи до лекарства на полке аптеки? На все эти вопросы методы химической информатики (хемоинформатики) помогают ответить прямо сейчас учёным по всему миру.

Ваша дорога будет полна открытий. Вы познакомитесь с методами анализа больших данных, научитесь работать с молекулярными структурами и моделями, поймете, как современные алгоритмы машинного обучения помогают выявлять закономерности в сложных химических системах. Эта книга создана для того, чтобы сделать эти знания доступными и понятными, даже если вы только начинаете свой путь в этой области.

Приготовьтесь к тому, что эта наука откроет перед вами новые горизонты – от фундаментальных исследований до практических приложений в медицинской химии, экологии и материаловедении. Мы уверены, что полученные знания вдохновят вас на собственные открытия и помогут внести свой вклад в развитие науки и технологий будущего.

В представляемой монографии будут также разобраны ряд аспектов, связанных с вычислительной токсикологией, так как эти две области науки неразрывно связаны, образуя мощный междисциплинарный тандем, критически важный для современной оценки рисков химических веществ. *Химическая информатика* предоставляет необходимый набор инструментов для обработки и анализа данных и построения предсказательных моделей, без которых вычислительная токсикология просто не смогла бы существовать. *Вычислитель-*

ная токсикология, в свою очередь, использует этот инструментарий и данные, предоставляемые химической информатикой, для построения моделей, предсказывающих токсичность химических веществ. Эти модели варьируются от относительно простых моделей, связывающих структурные факторы молекул с конкретными видами и параметрами токсичности (например, острая токсичность, мутагенность, цитотоксичность и др.), до сложных системных биологических моделей, интегрирующих данные о взаимодействии веществ с биологическими мишенями (молекулярный докинг), путях метаболизма и сигнальных путях. Без точного цифрового представления молекул, эффективных способов сравнения структур и огромных баз данных о известных параметрах токсичности (курируемых методами химической информатики), обучение и валидация таких предсказательных моделей были бы невозможны.

Связь этих двух областей становится всё более важной по нескольким ключевым причинам:

1. **Сокращение испытаний на животных:** вычислительные методы, основанные на данных химической информатики, позволяют проводить первоначальную оценку токсичности *in silico* (на компьютере), значительно сокращая количество необходимых дорогостоящих и этически спорных экспериментов на животных. Это особенно важно в контексте требований национальных и международных регуляторов и растущего общественного спроса на альтернативы.

2. **Оценка безопасности огромного числа химических веществ:** человечество окружено колоссальным количеством химических веществ, число которых растет с каждым днем – промышленные химикаты, новые фармацевтические кандидаты, наноматериалы, продукты распада и переработки отходов и вещества из окружающей среды (список можно продолжить). Экспериментально протестировать все из них на все возможные виды токсичности просто нереально. Вычислительные подходы в токсикологии, питаемые данными хемоинформатики, позволяют проводить приоритизацию веществ для дальнейшего тестирования на основе предсказанного риска и скрининг огромных виртуальных библиотек соединений на ранних стадиях разработки (например, лекарств или пестицидов/гербицидов) на предмет потенциальной токсичности.

3. **Понимание механизмов токсичности:** интеграция данных о структуре, свойствах, взаимодействиях с мишенями и путях, управляемая методами химической информатики и сложным моделированием, позволяет глубже понять почему вещество токсично. Это знание критично для разработки более безопасных химикатов («зеленая химия»), прогнозирования токсичности для новых соединений (в том числе для тех, которые еще не синтезированы), оценки вероятности появления побочных эффектов у лекарств и пищевых добавок, а также оценки рисков для сложных смесей.

4. **Ускорение разработки и снижение затрат:** в фармацевтике и химической промышленности раннее выявление потенциальной токсичности с помощью вычислительных методов позволяет отсеять неперспективные кандидаты на самых ранних, наименее затратных стадиях разработки, экономя огромные ресурсы и ускоряя вывод на рынок безопасных продуктов.

Таким образом, химическая информатика является «двигателем», обеспечивающим вычислительную токсикологию топливом (данными) и инструментами (алгоритмами поиска и сравнения молекул, методами представления, предсказательными моделями и др.). Этот симбиоз лежит в основе революции в оценке безопасности, позволяя нам более эффективно, этично и глубоко понимать потенциальный вред химических веществ, что крайне важно для защиты здоровья человека и окружающей среды в мире, насыщенном синтетической химией. Развитие методов машинного обучения и искусственного интеллекта ещё больше усиливает возможности этого тандема, открывая новые горизонты в предсказательной токсикологии и анализе химических рисков.

Следует подчеркнуть, что данная монография по своему характеру является *введением* в эту бурно развивающуюся область с целью познакомить читателя с основными понятиями и концепциями химической информатики и возможностями соответствующих инструментов в области вычислительной токсикологии и других важных отраслях. Огромная широта, глубина и скорость развития каждой из обсуждаемых подтем – от специфических алгоритмов генерации молекулярных дескрипторов до сложных системных биологических моделей токсичности – делают невозможным их исчерпывающее рассмотрение в рамках одной книги.

Ежегодно публикуются десятки тысяч статей, посвященных новым алгоритмам представления молекул, методам машинного обучения для прогноза токсичности, интеграции молекулярных данных, созданию огромных баз знаний и внедрению этих подходов в регуляторную практику и промышленность. Этот взрывной рост – ответ на обсуждаемую выше острую потребность в эффективных, этичных (сокращающих испытания на животных) и экономичных способах оценки рисков для здоровья человека и окружающей среды от бесчисленного множества существующих и новых химических веществ.

Именно здесь возникает ключевая проблема, которую призвана решить данная книга: как сориентироваться новичку или специалисту из смежной области в этом океане информации? Действительно, эта область глубоко междисциплинарная и объединяя химию, биологию, компьютерные науки, математическую статистику, машинное обучение и науку о данных. Ее терминология, методы и источники данных могут показаться сложными и фрагментированными как для начинающих исследователей, так и для экспертов из смежных отраслей. Поэтому мы постараемся дать ответы на следующие вопросы:

- Как фундаментальные концепции химической информатики конкретно питают и делают возможными сложные модели вычислительной токсикологии?
- Каковы основные «кирпичики» и логика построения систем оценки токсичности и других свойств молекул *in silico*?
- Как разные подходы (статистические модели «структура-свойство», молекулярное моделирование, более сложные системные модели) соотносятся друг с другом и решают различные задачи?
- Каковы реальные возможности и текущие ограничения этих методов?

Данная монография была написана для того, чтобы помочь читателю сформировать структурированное введение и «общую карту местности» в этой области.

Для более глубокого погружения в конкретные аспекты химической информатики и вычислительной токсикологии настоятельно рекомендуется обращаться к специализированным обзорным статьям, монографиям, научным базам данных и вычислительным платформам. По мере изложения материала будут предоставлены ссылки на ключевые публикации и фундаментальные источники, которые послужат отправной точкой для дальнейшего, более детального изучения этой критически важной и захватывающей области науки.

Добро пожаловать в мир химической информатики – мир, где химия встречается с цифровыми технологиями, а возможности ограничены только вашим воображением!

1.1. Определение и история химической информатики

Химическая информатика (хемоинформатика) – это быстро развивающаяся мультидисциплинарная область, которая объединяет в себе химические науки, информатику, биологию и математику для решения сложных задач, связанных с анализом и обработкой химических данных [1]. За последние двадцать с небольшим лет хемоинформатика значительно изменила подходы к проведению химических исследований, обеспечив доступ к информации в масштабах, недостижимых для традиционных методов. В этом контексте важно понять, как именно хемоинформатика влияет на различные аспекты химии и смежных с ней наук [2]. Само определение хемоинформатики до сих пор не формулируется лишь единственным образом, поскольку эта дисциплина находится на стыке многих смежных научных областей. В разное время учёные с большим авторитетом давали определения предмету, исходя из трендов времени, своего опыта и потребностей тех научных задач, которые они решали. Таким образом, в 1998 году, одним из первых, кто сформулировал определение хемоинформатики, вошедшим в обиход, был доктор Фрэнк Браун. В одной из своих статей он писал: «Использование информационных технологий и менеджмента стало ключевой стадией в открытии лекарств. Хемоинформатика – это объединение таких информационных ресурсов для преобразования данных в информацию и информации в знания с целью более быстрого и точного принятия решений в области поиска и оптимизации лекарственных средств» [3].

Спустя некоторое время в 2003 году в своей книге «Handbook of Chemoinformatics» Иоганн Гастайгер дал более широкое определение, которое вполне актуально по сей день: «Хемоинформатика – это наука, использующая методы информатики для решения химических проблем» [4]. Книга не теряет своей актуальности в настоящее время и может быть полезна как студентам, только начинающим свой путь в хемоинформатике, так и профессионалам в области, поскольку предоставляет всесторонний обзор методов и приложений химической информатики.

Необходимость возникновения такого научного направления стала очевидна ещё в начале XX века. Экспоненциальный рост новых химических соединений, их свойств и способов синтеза выявил запрос на хранение огромного количества химической информации в единых базах данных. В первую очередь актуальность цифровых баз данных понимали университеты и фармацевтические компании поскольку большое количество исследований даже в рамках одного проекта, информацию о которых нужно где-то хранить, могло превышать десятки тысяч химических соединений разных классов, включая и их физико-химические свойства [5].

Таким образом, в 1907 году была основана одна из первых баз данных с научной литературой по химии CAS (Chemical Abstract Service). Однако ее компьютеризированная версия появилась только в 1965 году. Изначально CAS была создана как часть Американского химического общества (ACS) с целью помочь учёным получать доступ к опубликованным научным работам. В 1965 году CAS разработала CAS Chemical Registry System, т.е. систему, которая использует уникальные номера CAS для идентификации химических веществ. Это стало важным шагом в стандартизации и упрощении поиска химической информации. И уже в 1980 году был запущен первый онлайн-сервис CAS Online, который позволял пользователям искать информацию о соединениях через специализированные терминалы [6]. На сегодняшний день база данных CAS содержит информацию о более чем 210 миллионах органических и неорганических соединений, а также около 75 миллионов последовательностей белков и нуклеиновых кислот. CAS продолжает обновлять свои базы данных, добавляя новые исследования и публикации [7]. Однако сейчас это далеко не единственная база данных, которая содержит в себе достаточное количество информации для анализа химических данных. В разделе (2.2) будут подробно описаны наиболее популярные и актуальные в настоящее время базы данных, которыми пользуются большинство химиков и не только. Также стоит отметить, что среди них будет освещена отечественная онлайн платформа ИИ с большой встроенной базой данных различных химических соединений под названием Синтелли.

После возникновения цифровых баз данных и разработки различных способов хранения / передачи химической информации в электронном формате, стали развиваться методы хемоинформатики, связанные с прогнозированием свойств химический соединений, которые ещё не были синтезированы, для оценки их потенциальной лекарственной пригодности или других важных свойств, полезных в смежных областях химии [8]. В числе таких свойств особое место занимает прогноз токсичности, поскольку ее оценка является важной частью для обеспечения безопасности новых химических соединений и их применения в различных отраслях, включая фармацевтику, сельское хозяйство, промышленность и военную отрасль.

С этого момента и начинается современная история хемоинформатики, так как подходы в прогнозировании этих свойств с каждым годом становятся всё более точными и эффективными благодаря развитию вычислительных технологий и алгоритмов машинного обучения.

1.2. Значение химической информатики в современной науке

Переоценить значение хемоинформатики в современной науке довольно сложно. Эта область повлияла на множество подходов, связанных с анализом большого количества химических структур совершенно различного состава – начиная от неорганических соединений в материаловедении и заканчивая сложными органическими макромолекулами в биологических системах. Таким образом, ее применение на данный момент не ограничивается только фармацевтической отраслью, поскольку и в других направлениях химии существует потребность в прогнозе тех или иных свойств молекул, их структуры и взаимосвязей между ними. К примеру, в материаловедении требуется разработка новых материалов с заданными свойствами (катализаторы для различных реакций) [9], в экологии и других областях необходима предварительная оценка токсичности химических соединений (наличие опасных для здоровья человека соединений в воде) [10], а в фармацевтике присутствует постоянный поиск биологически активных лигандов с целью их дальнейшего преобразования до полноценных лекарств.

Наибольшее применение хемоинформатика нашла именно в фармацевтической отрасли [11]. Ведь там хоть и сложный, но достаточно понятный процесс монетизации тех идей, которые модели, построенные на химических данных, могут привнести. Тем самым, они дают возможность специалистам быстрее создавать конечный продукт. В настоящее время в любой большой фармацевтической компании присутствует научный отдел, занимающийся предварительным поиском потенциально активных веществ среди огромного количества соединений с помощью методов химической информатики. Подобно тому, как кладоискатель ищет драгоценные металлы среди необъятного пространства нашей планеты, так и исследователи «охотятся» за уникальными химическими структурами, способными проявлять биологическую активность по отношению к определенной мишени.

Почему же это такая сложная задача? Всё дело в том, что химическое пространство насчитывает порядка 10^{60} малых (масса до 500 Да) синтетически доступных химических соединений, которые удовлетворяют общим правилам лекарственной подобности (правило пяти Липински). Это сравнимо со всем количеством атомов во вселенной. В этом числе всего лишь 10^8 известных человечеству химических структур на данных момент [12]. А ко всему прочему, 10^4 новых соединений появляется каждый день. С таким объемом данных не справилась бы ни одна научная группа, работая в отсутствии методов машинного обучения (которые стали полноценной частью хемоинформатики) или просто цифровой химической информации. Именно в связи с этим методы хемоинформатики плотно вошли в индустрию создания лекарств, с целью помочь человечеству среди огромного количества информации находить те «золотые» структуры, способные помочь в борьбе с неизлечимыми на данный момент болезнями, такими как рак, Альцгеймер, ВИЧ и многие другие [13].

Влияние хемоинформатики велико не только в коммерческой отрасли, но также и в научной среде. Стоит упомянуть, что в 2024 году Нобелевскую премию по химии дали за прогнозирование трехмерной структуры белков с помощью методов машинного обучения. Дэвид Бэйкер, Демис Хассабис и Джон Джампери создали алгоритм AlphaFold, способный из аминокислотных последовательностей воспроизводить трехмерную структуру белка [14]. Это стало настоящим прорывом с точки зрения подхода к прогнозированию роли биологических мишеней в организме человека. Ведь если мы знаем их строение, то можем ответить на вопрос о том, какую функцию они выполняют: являются переносчиками веществ в организме; ферментами, катализирующими различные реакции, и т.д. Данная премия вызвала много споров и негодования среди научного сообщества, поскольку ставит под сомнение наличие «чистой» химии в таком исследовании. Однако это новая реальность, которую всем стоит принять и осознать, что хемоинформатика и искусственный интеллект открывают новые горизонты для понимания сложных биологических систем. Эти технологии не только ускоряют исследования, но и позволяют делать более точные предсказания, что может привести к революционным открытиям в области медицины и биохимии. Важно адаптироваться к этим изменениям и использовать их потенциал для решения актуальных проблем, стоящих перед человечеством.

1.3. Основные области применения химической информатики

Продолжая мысль о применении хемоинформатики, нельзя не отметить, что ее главной задачей является прогнозирование свойств химических соединений. Такими свойствами могут быть любые показатели, начиная от температуры кипения, заканчивая ролью соединения в механизме химической реакции. Вот лишь небольшой список самых популярных свойств, учитываемых при разработке новых лекарственных препаратов:

1) Растворимость – определяет, насколько хорошо препарат может растворяться в воде или биологических жидкостях.

2) Липофильность – влияет на способность молекулы проникать через клеточные мембраны и распределяться в организме.

3) Токсичность – оценка потенциального вредного воздействия соединения на клетки и ткани, что позволяет избежать разработки опасных препаратов.

4) Сложность синтеза – определяет возможность масштабирования синтеза конкретного соединения в промышленных масштабах или сложность создания молекулы в лабораторных условиях.

5) Экологические свойства – предсказывают насколько реагенты или продукты разложения могут быть опасны для окружающей среды.

6) Биологическая активность – оценка эффективности воздействия препарата на определенную мишень в организме.

Это показатели, которые обязательно должны быть учтены для успешного создания соединений, способных стать безопасными лекарственными средствами.

Кроме того, рассматривая методы молекулярного моделирования, стоит сказать о возможности прогнозирования пространственного взаимодействия лиганда (малой молекулы) с его мишенью, с оценкой качества связывания с помощью специальных оценочных функций (алгоритмов расчёта комплементарности взаимодействий). Такой процесс называют «докингом». Его целью является предсказание благоприятной геометрии, при которой лиганд связывается со своей биологической мишенью (рис 1).

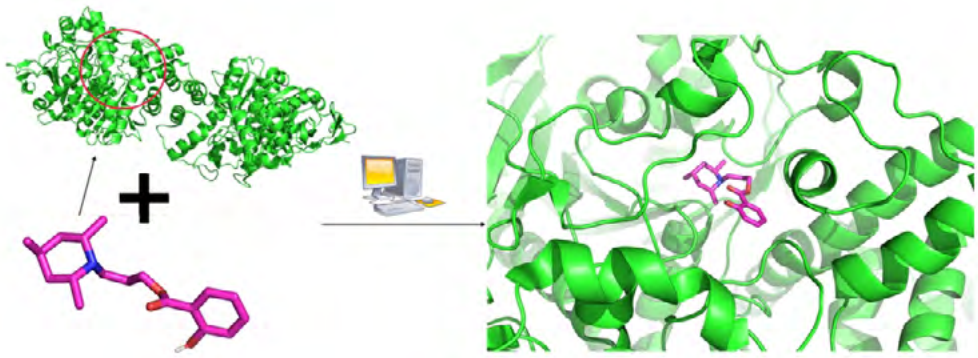


Рис. 1. Схематичная визуализация стыковки молекулы с биологической мишенью

Этот подход позволяет исследователям эффективно отбирать вещества-кандидаты для дальнейших экспериментов, минимизируя время и ресурсы, затрачиваемые на синтез и тестирование соединений *in vitro*. Докинг также помогает в понимании механизмов взаимодействия между лигандами и мишенями, что может способствовать разработке более эффективных и селективных лекарств [15].

Перечислять способы применения хемоинформатики можно очень долго, поскольку палитра ее возможностей с каждым годом только растет. В последующих главах этой книги Вы найдете концептуальные объяснения различных методов, которые используются в современных исследованиях для достижения максимальной эффективности использования времени, денег и других важных ресурсов.

1.4. Основные концепции химической информатики

Говоря об основных концепциях хемоинформатики, стоит провести четкую границу между смежными с ней областями, такими как биоинформатика, вычислительная химия и квантовая химия. Различие между ними заключается во входных параметрах, которые берутся за основу для исследования. Хемоинформатика в первую очередь отталкивается от известных данных, которые могли быть получены из эксперимента (измерения log P, температуры кипения и т.д.) или сгенерированы на основе тех же хемоинформатических методов (отображение молекул в формате SMILES, InChi, и т.д.). Работа с такими данными

подразумевает построение предсказательной модели тех свойств, которые нас интересуют. А данные, с которыми приходится работать, могут быть совершенно различными как по способу их генерации, так и по физико-химическому смыслу [16].

Биоинформатика в свою очередь использует дискретные коды, такие как аминокислотная последовательность, благодаря которым выполняет различные модификации с биополимерами. Преимущество таких данных заключается в понятном конечном наборе видов данных (20 аминокислот) и однозначной теории кодирования для работы с ними [17].

Вычислительная химия изначально отталкивается от математической модели, описывающей молекулу. В этом моменте стоит однозначно определить вид теории, в которой мы описываем молекулу. В терминах молекулярной механики молекула рассматривается как совокупность шариков (атомов) и пружинок (связей). Эта модель позволяет описывать молекулы с использованием классических физических законов, что делает расчёты относительно быстрыми и менее ресурсоемкими по сравнению с квантово-химическими методами. Однако, несмотря на свою простоту, молекулярная механика имеет свои ограничения: она не учитывает квантовые эффекты, которые могут быть критически важны для понимания поведения электронов в молекуле и протекания химических реакций [18].

Для более точного описания молекул и их взаимодействий используются квантово-химические методы, такие как метод Хартри-Фока или теория функционала плотности (DFT) [19]. Эти методы рассматривают молекулы на основе принципов квантовой механики, позволяя более точно предсказывать энергетические уровни электронов, геометрию и, как следствие, свойства соединений.

Таким образом, выбор подходящей теории для описания молекулы зависит от конкретной задачи и требуемой точности. На практике часто используются комбинации методов: сначала применяются молекулярно-механические расчёты для предварительной оптимизации структуры, а затем более точные квантово-химические методы для детального анализа определенных аспектов системы. Это позволяет эффективно использовать вычислительные ресурсы и получать надежные результаты в области изучения химических соединений.

2. ОСНОВЫ ЦИФРОВИЗАЦИИ МОЛЕКУЛЯРНОЙ ИНФОРМАЦИИ

Одна из первых и ключевых задач хемоинформатики заключалась в представлении химических соединений в формате, понятном для компьютеров. Начиная с XIX века, химики обменивались информацией с помощью брутто-формул различных соединений. Однако абсолютное большинство брутто-формул нельзя назвать однозначными, поскольку одной и той же записи может соответствовать несколько химических соединений, которые различаются по своим химическими свойствами и классам. В связи с этим стала логичной разработка структурных формул, которые были способны отобразить индивидуальные структурные особенности молекулы в 2D формате. В химической информатике происходит нечто похожее.

Создание цифровых баз данных потребовало разработки алгоритмов для однозначного и уникального кодирования химических структур. Сразу несколько учёных по всему миру стремились решить эту задачу. В первую очередь необходимо было определить размерность пространства, в котором будут отображаться молекулы. Каждому химику понятно, что даже нарисованная на листе бумаги молекула не является истинным отображением ее реальной геометрии, поскольку подавляющее большинство из них далеко не плоские, каковыми они предстают на бумаге. Таким образом, трехмерное пространство стало единственным способом максимально точного представления молекул, так как именно в таком формате можно учесть все пространственные взаимодействия между атомами, которые существенно влияют на химические свойства и реакционную способность соединений.

Но кодирование структур сразу в формате 3D было бы невозможным без поэтапного понимания предшествующих стадий с размерностями 1D и 2D. Таким образом, сначала появились линейные нотации (1D), представляющие молекулу в виде строки символов, что позволило удобно хранить и обмениваться такой информацией между пользователями виртуальных библиотек. Затем уже матричные и табличные форматы представления молекул подоспели на помощь. С их помощью открылась возможность к 2D и 3D представлениям, что обеспечило химикам большую точность воспроизведения геометрий молекул в цифровом формате.

Стоит уточнить, что в химии также присутствует понятие 4D размерности, суть которой заключается в представлении ансамбля конформаций, доступных для каждой молекулы и их изменений во времени. Однако эта часть требует более детального погружения в методы молекулярного моделирования.

На самом деле, в каждой размерности представление молекул по сей день имеет свои ограничения и определенные условия для корректного перевода той или иной структуры в нужный формат. Это связано с тем, что химическая отрасль не стоит на месте – каждый день появляются новые, структурно более сложные соединения, которые требуют обновлений или корректировок некоторых правил кодирования структур. В связи с этим, компетентные специалисты

уже разработали и продолжают разрабатывать множество форматов цифровых файлов, в которых можно хранить химическую информацию под различные типы задач хемоинформатики. В этом разделе мы подробнее поговорим о правилах кодирования структур, различных файлах для хранения химической информации и способах их применения.

2.1. Представление малых химических структур

В данном разделе под малыми химическими структурами мы будем понимать любые соединения, которые не превышают порог по молекулярной массе в 900 Дальтон. Заметим, что большинство биологических мишеней имеют массу, сильно превышающую это значение. К примеру белки, нуклеиновые кислоты, полисахариды и т.д. являются классическими представителями природных биополимеров с длинными цепочками, состоящими из определенных соединений, работа со структурой которых, будет рассмотрена в другом разделе.

Такое разделение химических соединений обусловлено разными нагрузками на ПК при обработке данных, а также способами их генерации и работе с ними.

2.1.1. Линейные представления

Линейные представления берут свое историческое начало ещё со времен использования брутто-формул, поскольку они, de facto, являются представлениями химической структуры в виде строки символов, состоящей из букв и цифр. В один ряд с брутто-формулами можно поставить и систематическое название молекулы, несмотря на то что количество используемых символов в такой строке сильно увеличивается из-за наличия скобок, запятых, обозначений стереохимии и других деталей, которых нет в названии брутто-формул. Связано это с правилами более подробного описания структуры соединения, к которому уже давно привыкли химики, но не компьютеры. Ведь чем больше символов тратится на запись одной структуры, тем больше времени потребуется вычислительной машине для ее обработки и дальнейших манипуляций с ней.

Все перечисленные недостатки классических способов записи молекул побудили экспертов разработать уникальный свод правил, по которым химическую структуру можно было бы кодировать:

1. Однозначно (одно название – одна структура).
2. Уникально (одна структура – одно название).
3. Быстро (скорость кодирования вручную).
4. Экономно (память, ресурсы ПК).
5. Обратимо (перевод структуры в код и обратно).

В результате проделанной работы одними из самых популярных линейных нотаций в свое время стали WLN (Wiswesser Line Notation) [20], SMILES [21] и InChI [22]. Последние две не теряют своей актуальности по сей день и широко применяются в различных химических программных пакетах. Об их правилах, преимуществах, недостатках мы и поговорим подробнее.

2.1.1.1. Линейная нотация Висвессера (WLN)

При перечислении главных по значению линейных нотаций не случайно WLN часто ставят на первое место. Нет, далеко не потому, что это самый популярный способ кодирования молекул на данный момент, скорее наоборот. Этот вариант перевода химических соединений в строку был разработан ещё в конце 1940х – начале 1950х годов Уильямом Джозефом Висвессером - американским химиком, работавшим в штате Пенсильвания [20]. Он был первым, кому удалось сформулировать понятный свод правил, с помощью которого человек и компьютер могли обмениваться химической информацией. И как часто случается с первопроходцами, в настоящее время этот способ практически полностью утратил актуальность из-за своей специфичности и сложного согласования с другими методами. Однако он по праву считается родоначальником линейных нотаций, поскольку создал серьезный фундамент для других методов, таких как SMILES и InChI. WLN продемонстрировала важность стандартизированного подхода к представлению химических структур, что стало основой для дальнейших разработок в области хемоинформатики. Таким образом, хотя WLN и считается устаревшей на сегодняшний день, ее вклад в развитие линейных нотаций остается неоспоримым и служит важным напоминанием о том, как эволюционировали методы представления химической информации.

Учитывая серьезное влияние алгоритмов, присутствующих в линейных нотациях Висвессера, стоит кратко рассмотреть свод этих правил, по которым кодировались химические структуры.

Основой для представления молекулы в WLN является процедура разделения ее на фрагменты, которые имеют свои идентификаторы в алфавитной системе метода. Различные фрагменты могут быть представлены в виде определенных букв или цифр, перечисление которых в строгом порядке соответствует связям этих фрагментов в молекуле.

Всего в нотации WLN присутствует 40 символов:

1) Буквы от «A» до «Z» – 26 символов, которые используют для обозначения специальных структурных фрагментов в молекуле.

2) Цифры от 0 до 9 – 10 символов, которые используют для обозначения количества углерода в алкильных фрагментах.

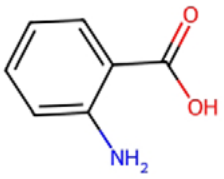
3) А также 4 символа: «&», «/», «-» и пробел. Одни отвечают за разветвления в молекуле («&», «/»), а оставшиеся («-») и (пробел) помогают в разделении структурных фрагментов.

В случае с атомами, проще всего указать водород, поскольку он обозначается привычным символом H. Однако с углеродом, азотом и кислородом дело обстоит сложнее, поскольку для каждого из них возможны различные виды связей и их количество. Кроме того, для каждой функциональной группы (амин, имин, гидроксил, карбонил и т.д.) есть свой уникальный символ, который в строгой последовательности кодирует ее содержание в химическом соединении. Ароматическим и просто циклическим фрагментам также соответствуют специальные обозначения в виде латинских букв. Уже в нескольких методических пособиях на русском языке подробно рассматривались эти правила коди-

ровки в нотации WLN. Если вам хочется разобраться с этим процессом подробно, рекомендуем к прочтению учебное пособие от Уральского федерального университета «Компьютерное представление химической информации: учебное пособие» [23]. Здесь мы лишь приведём несколько примеров кодирования с помощью этой нотации и дадим небольшие пояснения для общей оценки такого подхода в представлении химических структур.

Ниже приведены несколько примеров названий молекул, которые достаточно полно отображают возможности кодировки структур с помощью WLN.

Табл. 1. Примеры кодирования различных молекул в WLN

		
а) WLN нотация: 1V1	б) WLN нотация: ZR BVQ	в) WLN нотация: 1U4UU

В случае с ацетоном (а), цифра 1 обозначает метильную группу (-CH_3) с одной и другой стороны от карбонильной группы, которая по правилам WLN нотации записывается символом «V». Во втором примере (б) продемонстрирована возможность кодирования аминогруппы (-NH_2) символом «Z», которая присоединена к фенильному ароматическому кольцу - символ «R», в котором, в свою очередь, содержится карбоксильная группа (-COOH), кодируемая символами «VQ» (V – карбонил, Q – гидроксил). Символ «B» в данном случае обозначает наличие этой самой карбоксильной группы, как заместителя в бензольном кольце. Пример (в) демонстрирует, что тройная связь в алкинах (в конце) кодируется с помощью двух букв «UU», двойная связь в алкенах просто буквой «U», а цифры 1 и 4 показывают количество углеродных атомов по обе стороны от двойной связи.

Подводя итоги и давая оценку WLN, стоит сказать, что очевидными преимуществами этой нотации являются:

1) краткость записи;
2) однозначность;
3) уникальность.

Почему же тогда в настоящее время этот способ практически неприменим? WLN содержит в себе довольно много правил, при учете которых вручную легко сделать ошибку. Кроме того, сегодня, при большом изобилии форматов для хранения файлов с химической информацией, важно уметь быстро переводить молекулу из одного формата в другой. Каждый формат создавался для работы в определенных коммерческих или некоммерческих программных пакетах. WLN очевидно имеет сложности при автоматической конвертации структур в различные представления для работы с разными форматами. Вполне возможно, это и были основные причины, по которым линейная нотация Висвессера вышла из обихода компьютерных химиков.

2.1.1.2. SMILES

Куда лучше с перечисленными выше задачами справилась линейная нотация SMILES (Simplified Molecular Input Line Entry System – система упрощенного представления молекул в строке ввода), изобретенная Дэвидом Вейнингером в конце 1980-х [21]. Она по сей день остается самой популярной методикой перевода химических структур в строку и используется во всех основных химических программных пакетах. Ее основным преимуществом по сравнению с WLN является наглядность для химиков, поскольку атомы и связи в названиях отражены аналогично тому, как они выглядят в структуре соединения.

Дэвид Вейнингер опубликовал статью, посвященную SMILES в 1988 году, много ссылаясь на работы своего предшественника Уильяма Джозефа Висвессера. Сам он писал: «SMILES основана на принципах теории молекулярных графов и позволяет строго специфицировать структуры, используя очень маленькую и естественную грамматику. Система нотаций SMILES идеально подходит для высокоскоростной машинной обработки. Благодаря простоте использования химиком и совместимости с машинами, можно разработать множество высокоэффективных химических компьютерных приложений, включая генерацию уникальной нотации, поиск по базе данных с постоянной скоростью, гибкий поиск подструктур и модели предсказания свойств» [21].

С учетом всего вышесказанного очевидно, что подробное знание правил генерации и понимание названий SMILES просто необходимо не только современному хемоинформатику, но и в том числе химику. Именно поэтому ниже будут рассмотрены правила, по которым происходит кодирование различных соединений в строки.

Начать стоит с обозначений атомов (табл. 2). Каждый атом кодируется своим атомным символом в квадратных скобках, однако некоторые (C, P, O, Cl, S, N) могут быть записаны и без них. В свою очередь атомы водорода могут не указываться вовсе (водородно-супрессивные графы) или же указываться (водородно-полные графы) по желанию составителя. В случае отсутствия атомов водорода в закодированной структуре при переводе они отображаются автоматически с учетом валентности атомов, к которым относятся.

Табл. 2. Обозначение атомов в SMILES

Обозначение	**Струкутра**	**Название**
C	CH_4	Метан
P	PH_3	Фосфин
Cl	HCl	Соляная кислота
O	H_2O	Вода
S	SH_2	Сероводородная кислота
N	NH_3	Аммиак
[Au]	Au	Атомарное золото

Фрагменты, имеющие заряд, записываются в квадратных скобках вместе с количеством атомов водорода и типом заряда (табл. 3). Обратите внимание на числа, стоящие перед знаком заряда и после него. Если число стоит перед знаком плюса/минуса, то оно обозначает количество водородов в структуре. Если же после знака, тогда оно относится к количественной характеристике самого заряда.

Табл. 3. Обозначение заряженных группировок в SMILES

Обозначение	Название
[H+]	Протон
[OH-]	Гидроксил анион
[OH3+]	Гидроксоний
[Fe+3]	Катион железа (III)
[NH4+]	Катион аммония

Кроме того, в SMILES считаются синонимичными записи [Fe+3] и [Fe+++], что говорит об алгоритмической гибкости представлений такого рода фрагментов.

После того, как определились с атомами, стоит задать вопрос о типе их связанности друг с другом. Обозначения связей в данной нотации также максимально сопоставимы с химическими традициями (табл. 4):

1) Одинарная связь либо обозначается символом «-», либо опускается по умолчанию.
2) Двойная связь обязательно обозначается символом «=».
3) Тройная же связь имеет формат записи через решётку «#» (также обязательно указывать при наличии).

Табл. 4. Обозначение связей в SMILES

Пример кодирования	Интерпретация
CC	CH_3-CH_3
C=C	CH_2=CH_2
COC	CH_3-O-CH_3
CCO	CH_3-CH_2-OH
O=C=O	CO_2
O=CO	HCOOH
C#N	HCN
[H][H]	H_2

Для того, чтобы избежать путаницы в расстановке водородов в конечной структуре, стоит задавать себе два простых вопроса на каждом атоме. Сколько связей от него уже отходит? Какая вообще у него валентность (по умолчанию принимается наименьшая из возможных)? Разница между вторым и первым числом всегда даст вам ответ на вопрос о количестве атомов водорода у каждого элемента в структуре.

Внимательный читатель уже на этой стадии обнаружит, что SMILES не совсем соответствуют требованию об уникальности названий. К примеру, 6-гидрокси-1,4-гексадиен может быть записан тремя разными способами (табл. 5).

Табл. 5. Примеры кодирования молекулы 6-гидрокси-1,4-гексадиена в SMILES

Структура	Подходящее кодирование
CH2=CH-CH2-CH=CH-CH2-OH	C=CCC=CCO
	C=C-C-C=C-C-O
	OCC=CCC=C

Для того, чтобы это исправить, в SMILES добавили концепцию канонической записи (Canonical SMILES), которая предназначена для обеспечения уникальности представления молекул. Эта версия спецификации включает правила, позволяющие однозначно записать формулу молекулы, выбирая определенный порядок обхода атомов и связей. Это значит, что для одной и той же молекулы всегда будет генерироваться одна и та же каноническая запись, что исключает вероятность неуникальности названия.

После того, как стало понятно, каким образом записываются неразветвленные структуры, стоит перейти к разветвленным. Идея SMILES заключается в том, чтобы записывать любые ответвления структуры химического соединения в круглых скобках. То есть атомы, которые не входят в состав основной углеродной цепи, берутся в эти самые скобки. К примеру, два атома углерода в триэтиламине, соответствующие боковой цепи, будут записаны под одной круглой скобкой (рис. 2). В свою очередь следующий за закрытой скобкой углерод связан с азотом, перед которым эти самые скобки открывались.

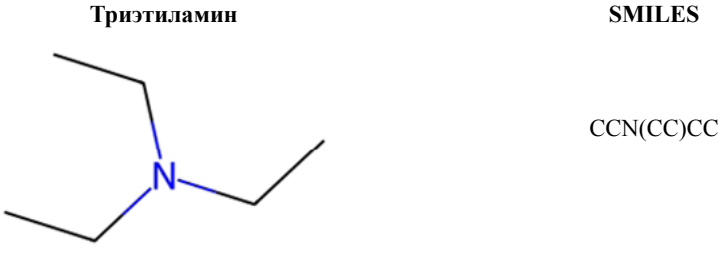
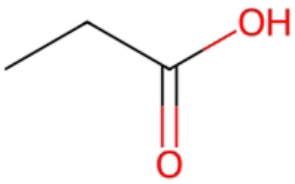


Рис. 2. 2D структура триэтиламина и его SMILES нотация

Если же ветвление сопровождается наличием кратной связи, она обозначается внутри этих скобок перед атомом. На примере пропионовой кислоты (рис. 3) можно убедиться в том, что кислород карбоксильной группы, который считается ответвлением от основной цепи, записывается согласно этим правилам в скобках с двойной связью перед ним. В свою очередь гидроксильная группа отображается атомом кислорода в контексте основной цепочки.

Пропионовая кислота **SMILES**

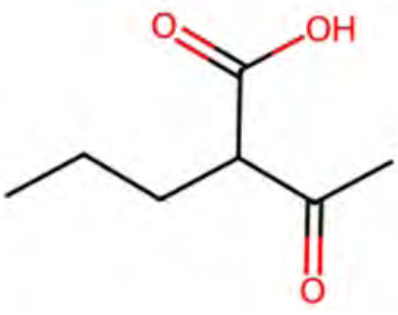


CCC(=O)O

Рис. 3. 2D структура пропионовой кислоты и ее SMILES нотация

Также, если присутствует ещё одно ответвление в боковой цепи, ставят дополнительные скобки в уже открытых первой боковой цепочкой. Для примера рассмотрим молекулу 2-ацетилпентановой кислоты. Стоит заметить, что ацетиловый фрагмент, который и так записан в скобках как боковой остаток, имеет метиловую группу, отображающуюся как дополнительное ответвление.

2-ацетилпентановая кислота **SMILES**



CCCC(C(C)=O)C(=O)O

Рис. 4. 2D структура 2-ацетилпентановой кислоты и ее SMILES нотация

Огромное значение имеет обозначение циклических структур. Для их кодирования существуют специальные правила, которые требуют условного разрыва одной из связей цикла. Механизм выглядит следующим образом:

1) Выбираем любую связь в цикле и формально разрываем ее.
2) Атомы, стоящие на концах оборванной связи, нумеруются одной и той же цифрой, которая при кодировании записывается после соответствующего атома.
3) Если циклов несколько, атомы у разорванных связей имеют разные номера, соответствующие номерам циклов в исходной молекуле.

На примере с циклогексаном можно убедиться в простоте данного механизма.

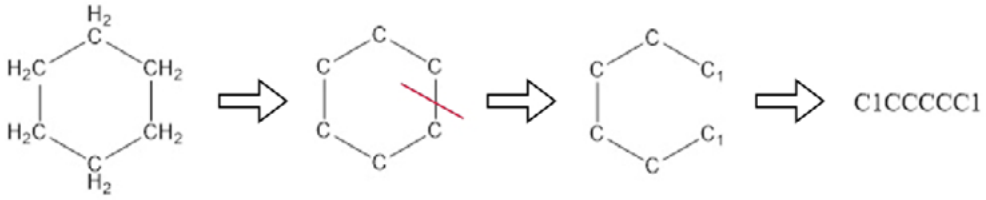


Рис. 5. Разрыв связи в циклогексане и его SMILES нотация

В случае с кубаном можно наблюдать разрыв нескольких связей и поочередную нумерацию каждого атома на конце соответствующих краев.

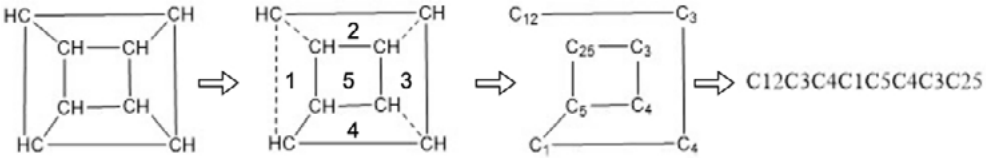


Рис. 6. Разрыв нескольких связей в кубане и его SMILES нотация

Стоит отметить, что повторное появление в названии структуры одной и той же цифры обозначает замыкание определенного цикла. Если при прочтении слева направо цикл открылся и успел закрыться, тогда эту же цифру возле атомов можно использовать повторно в следующем циклическом фрагменте. Для примера можно рассмотреть молекулу 1-(оксан-2-ил) пиперидина.

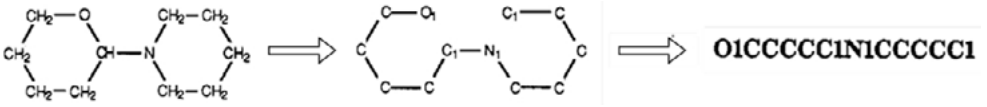


Рис. 7. Разрыв связей в 1-(оксан-2-ил)пиперидине ее SMILES нотация

Возможность повторного использования этих цифр позволяет задавать структуры с 10 и более циклами соответственно. Конструкции, в которых одновременно должно быть открыто более 10 циклов, встречаются крайне редко. Но даже в таком случае при необходимости можно указать числа больше 10, ставя перед двузначным числом знак процента. Например, углерод с кольцевыми замыканиями 2, 13 и 24 записывается как C2%13%24.

Очевидно, что большая часть различных циклических фрагментов представлена ароматичностью. Есть ли отличия при записи ароматических составляющих в молекуле?

Да, безусловно. Атомы, которые входят в состав таких фрагментов, записываются с маленькой буквы. Например, проследим отличия в записи циклогексана и бензола с пиридином.

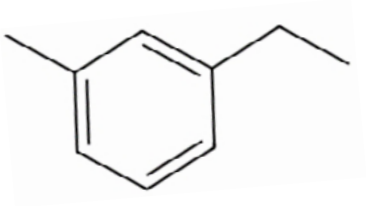
Циклогексан **Бензол** **Пиридин**

C1CCCCC1 c1ccccc1 c1ccccn1

Рис. 8. Структуры и SMILES нотации циклических не ароматических и ароматических молекул

Подключая уже известное нам правило о разветвленных структурах, можно без особых сложностей записывать ароматические соединения с их заместителями.

1-этил-3-метилбензол

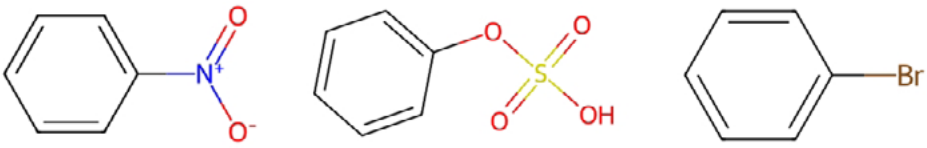


c1ccc(C)cc1CC

Рис. 9. 2D структура и пример полной SMILES нотации для 1-этил-3-метилбензола

Однако стоит ввести общепринятую запись нескольких популярных заместителей, характерных в первую очередь для ароматических систем, поскольку часто возникают споры, связанные с их строением.

Нитрогруппа **Сульфогруппа** **Галогеновые заместители**



а) c1ccccc1N(=O)=O
б) c1ccccc1[N+](=O)[O-]

c1ccccc1OS(=O)(=O)O

Brc1ccccc1

Рис. 10. Варианты записей в SMILES нотации нитрогруппы, сульфогруппы и галогеновых заместителей

В случае с нитрогруппой можно наблюдать две записи, которые отображают различные подходы обозначения связей, отходящих от азота. В записи под

буквой (а) подразумевается ковалентное связывание азота с двумя кислородами. А под буквой (б) учитываются делокализованные заряды по связи N – C. Для простоты восприятия всё чаще используют первый вариант записи нитрогруппы в SMILES.

Как можно было заметить, эта линейная нотация позволяет записывать заряженные атомы, между которыми существуют ионные или другие нековалентные взаимодействия. В таком случае элементы, не соединенные ковалентной связью, записываются через точку в любом порядке. К примеру, фенолят натрия будет иметь следующее кодирование:

Рис. 11. Варианты записей в SMILES нотации солей

Постепенно, вводя все больше новых правил, нельзя не заметить, что одну и ту же молекулу можно называть несколькими способами. Может меняться порядок атомов, с которых начинается кодирование. Нет однозначного механизма при выборе основной углеродной цепи и соответственно ее разветвлений. Даже обозначения некоторых функциональных групп (как мы успели убедиться) имеет двойственное представление.

Почему это проблема? При поиске целевых веществ в базах данных необходимо уникальное и однозначное соответствие химической структуры с ее названием. Это позволяет проводить максимально быстрый и удобный поиск среди миллионов молекул. Иногда, из-за неуникального кодирования, можно долго искать нужную структуру или вообще не найти абсолютно ничего, предварительно потратив много времени на поиск. Все эти проблемы побудили развитие SMILES до Canonical SMILES – аналогичную линейную нотацию, в которой были использованы дополнительные алгоритмы нумерации атомов и другие методы стандартизации, о которых мы упоминали выше. Подробнее о них можно почитать в работах Дэвида Вейнингера [24].

Вообще SMILES содержит довольно обширный и часто интуитивно понятный свод правил. К примеру, с их помощью также можно указывать стереохимические особенности молекулы (цис-транс, асимметрический атом) или же записывать полноценные реакции с помощью расширения SMIRKS. Об этих и других деталях SMILES Вы можете прочитать в 1 томе серии книг Т.И. Маджидова и коллег «Введение в хемоинформатику. Представления химических объектов» [25].

2.1.2. Графовые представления

Успешно освоив методы линейных нотаций и сумев представить молекулу в 1D формате, стоит перейти к 2D отображению структур. Конечно же, ничего не заменит обычной структурной формулы, которую химики рисуют на бумаге. Такая форма представления помогает им достаточно точно охарактеризовать молекулу, поскольку способна отобразить и различные типы атомов, и кратность связей, и даже некоторые геометрические особенности, связанные с цис-транс изомерией или хиральностью атомов. Однако «скормить» компьютеру рисунок с листа бумаги в чистом виде не кажется рациональным способом, поскольку дальнейшая обработка и хранение такого типа данных будет отнимать слишком много ресурсов и времени. Наиболее близким по смыслу способом представления структурных химических формул, который обладает математическим аппаратом и потенциалом к оцифровке, является граф.

Простым графом G (V, E) называется совокупность двух множеств – непустого множества V и множества E – множества неупорядоченных пар различных элементов множества V. Множество V называется множеством вершин, множество E называется множеством ребер [26]. Если упрощать фундаментальное определение графа (хотя бы количеством слов «множество») и формулировать его применительно к отображению химических формул, то стоит выделить две главные составляющие. Молекулярный граф – это совокупность вершин и соединяющих их ребер, где роль вершин играют атомы, а ребер - связи между ними. В своей сути, одно ребро – одна пара атомов, поэтому и говорят о том, что множество ребер – это множество пар вершин графа.

Такое представление молекулы несомненно выигрышно с точки зрения его оцифровки, поскольку любой молекулярный граф можно представить в виде матрицы, которую легко будет считывать и хранить в памяти сам компьютер. Однако с точки зрения реального представления структурных свойств молекулы, граф может похвастаться лишь отображением связности. Если граф непомеченный, что и происходит чаще всего, то химик лишается информации о типе атомов и кратности связей между ними, потому что все вершины и ребра в таком графе будут идентичны. Поскольку теория графов – это целый раздел математики со своим набором терминов и законов, стоит ввести несколько базовых определений и провести параллели с их интерпретацией под наши задачи.

Во-первых, графы делятся на ориентированные и неориентированные (рис 12).

Здесь всё просто, у ориентированных графов ребрам присвоено направление, а называют эти ребра дугами. В свою очередь, если ни одно из ребер графа не имеет направления, он называется неориентированным. Если **каждое** ребро графа имеет свое направление, то такой граф можно назвать турниром. Преимущество этих графов заключается в отображении вида связей между атомами. К примеру, таким способ можно показать донорно-акцепторную связь между двумя атомами в молекуле.

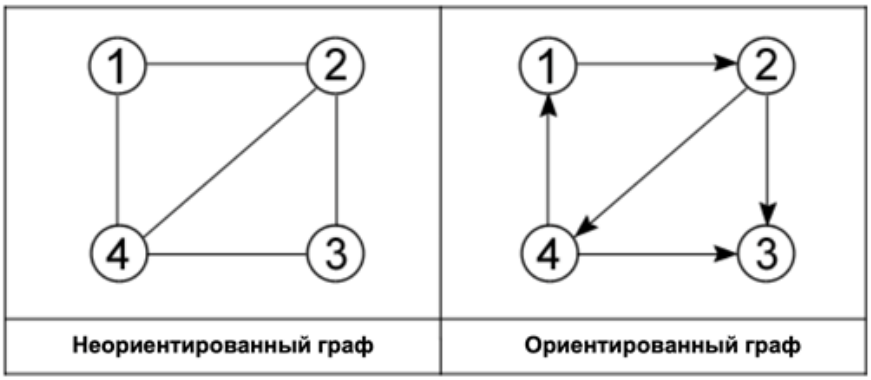


Рис. 12. Примеры ориентированного и неориентированного графа

Конечно же, традиционное представление молекулы – это неориентированный граф, узлы и ребра которого соответствуют тяжелым атомам и их связям в молекуле соответственно. Водороды в таких графах не указываются, поскольку могут быть расставлены, исходя из правил валентности (рис 13).

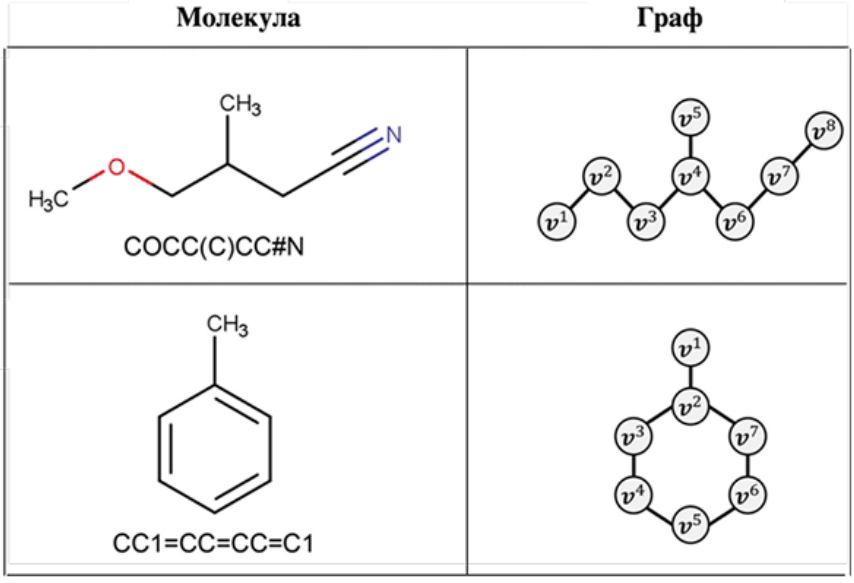


Рис. 13. Структурные формулы, SMILES и графовое отображение молекул 4-метокси-3-метилбутаннитрила и толуола

Возможно ли учесть типы атомов? Да, есть определенные виды графов, называемые помеченными. Это графы, в которых каждой вершине или ребру присвоена уникальная метка (число или символ из какого-нибудь алфавита). Это как раз и позволяет различать вершины и рёбра. Основоположником такой модификации выступил Алекс Роза в своей статье в журнале Theory of Graphs

в 1967 году. Там он представил несколько вариантов разметок графа, одну из которых впоследствии назвали Грациозной [27]. Граф называется грациозным, если его вершины помечены числами от 0 до $|E|$, и эта разметка порождает реберную разметку от 1 до $|E|$. При этом метка любого ребра e равна положительной разности между двумя вершинами ребра e (рис 14).

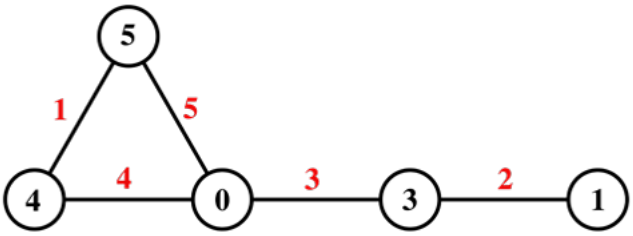
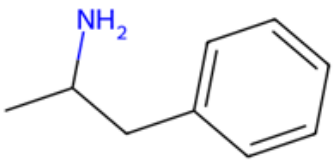


Рис. 14. Пример грациозного графа

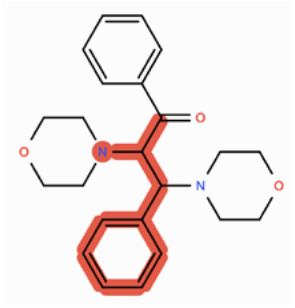
Таким образом в молекулярном графе при необходимости может содержаться информация не только о типе атомов, но и о кратности связей, которые также важно учитывать при исследовании.

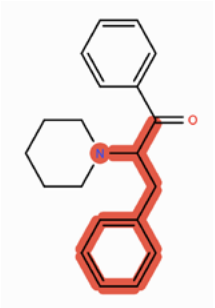
Возможность графового представления открывает перед химиками горизонт обмена информацией уже не только в формате линейной строки (1D), но и в 2D визуализации. С развитием теории графов стали доступны многие процессы, связанные с поиском структур по химическим базам. Например, молекулярные графы активно используются в подструктурном поиске, при котором необходимо найти молекулы, содержащие нужный структурный фрагмент «родительского» соединения (табл. 6).

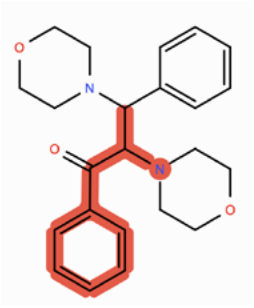
Табл. 6. Молекула амфетамина и содержащие в себе ее структуру соединения



Амфетамин – синтетический стимулятор центральной нервной системы







Для такого, как кажется, несложного действия на самом деле используется серьезный математический аппарат. Грамотная реализация алгоритмов основывается на задаче о поиске изоморфного подграфа. На этом этапе стоит ввести несколько важных терминов.

Некоторый граф называется подграфом основного, если множества вершин и ребер первого, являются подмножествами вершин и ребер второго соответственно. В свою очередь изоморфизм графов – это концепция, которая описывает отношение между двумя графами, когда они имеют одинаковую структуру, но могут отличаться в представлении. Выражаясь проще, два графа называются изоморфными, если существует взаимно однозначное соответствие между их вершинами, которое сохраняет структуру рёбер. Это означает, что можно «перекрасить» один граф в другой, не изменяя его топологию.

Однако строгое определение сложного термина никогда не помешает, поэтому следующий абзац для сильных духом читателей.

Графы $G_1(V_1, E_1)$ и $G_2(V_2, E_2)$ изоморфны, если существует биективная функция h: $V_1 \to V_2$, сохраняющая смежность:

$$\forall\ v \in V_1\ \exists u = h(v) \in V_2 :\forall\ v_1 \neq v_2\ h(v_1) \neq h(v_2),$$

$$e_1 = (u,v) \in E_1 \Rightarrow e_2 = (h(u),h(v)) \in E_2,$$

$$e_2 = (u,v) \in E_2 \Rightarrow e_1 = (h^{-1}(u),h^{-1}(v)) \in E_1.$$

Изоморфизм графов есть отношение эквивалентности. Графы рассматриваются с точностью до изоморфизма, то есть рассматриваются классы эквивалентности по отношению изоморфизма. Чтобы доказать, что два графа изоморфны, достаточно найти функцию $h : V_1 \to V_2$ из определения [26].

В качестве примера на рис (15) приведены изоморфные графы. Посмотрев на них, можно заметить, что все они имеют одинаковое количество вершин и ребер, а также количество вершин с одной и той же валентностью.

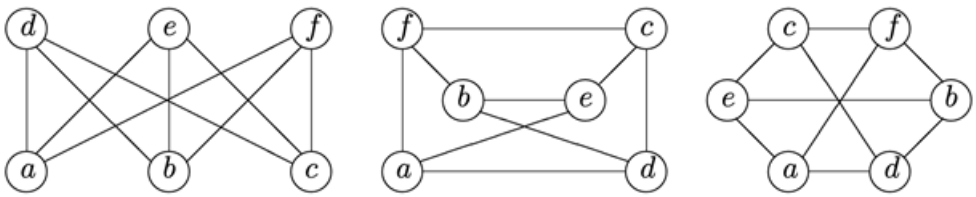


Рис. 15. Примеры изоморфных графов [26]

Возвращаясь к подструктурному поиску, скажем, что задача поиска изоморфного подграфа – это вычислительная задача, в которой входом являются два графа G и H. А далее нужно определить, не содержит ли G подграф, изоморфный графу H. Поскольку введенные нами термины подсветили глубину смыслов, заложенных в эту задачу, становится очевидным, что без хорошего математического аппарата здесь не обойтись. И в 1976 году Джулианом Ульманом был предложен одноименный алгоритм, позволяющий за полиномиальное

время вычислять наличие изоморфного подграфа в исходных. Подробнее о нем можно прочитать в статье от 2010 года, т.к. он был значительно изменен с момента его изобретения [28].

Завершая тему графовых представлений, стоит упомянуть, что конечный формат хранения данных такого типа в компьютере в основном представлен различными матрицами. Наиболее распространённые из них – это матрицы смежности и матрицы инцидентности. Матрицы смежности – это квадратные матрицы, в ячейках которых могут стоять 0 или 1, отображающие наличие связей между двумя атомами в графе молекулы (0 – если связи нет, 1 – если связь есть). Рассмотрим матрицу смежности метана (рис 16). В ней показано, как связи между 5 атомами метана располагаются симметрично относительно главной диагонали матрицы, позволяя хранить информацию о ней в цифровом формате. Симметричность и наличие 0 на главных диагоналях в таких матрицах обусловлено тем, что атом не может быть соединен сам собой.

Индекс	1	2	3	4	5
1	0	1	1	1	1
2	1	0	0	0	0
3	1	0	0	0	0
4	1	0	0	0	0
5	1	0	0	0	0

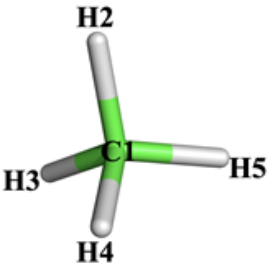


Рис. 16. Матрица смежности молекулы метана

2.1.3. Трехмерные представления (3D)

Было бы странно остановиться на двухмерном представлении молекул, понимая, что большинство из них имеют трехмерные структуры. Возможность учитывать пространственные особенности химических соединений позволили табличные форматы представления данных, в которые можно записывать координаты атомов по трем осям, а также другие признаки, включая типы связей и различные свойства молекул (хиральность, частичные заряды на атомах и т.д.). Наиболее распространенными табличными форматами в химической информатике являются MOL/MOL2 [29] и PDB [30]. Первые два чаще используются для отображения малых молекул, а вот в PDB кодируются белковые структуры, которые можно скачать из соответствующих баз данных. Рассмотрим организацию таких файлов.

Для того чтобы компьютеру понять, как атомы связаны между собой, используют матрицы смежности или их близкие аналоги. В форматах MOL и MOL2 после координат атомов реализованы специальные матрицы, позволяющие понять вид взаимосвязей между ними в молекулах. Рассмотрим организацию таких файлов на примере молекулы фосфорорганического отравляющего вещества Зарина (рис 17).

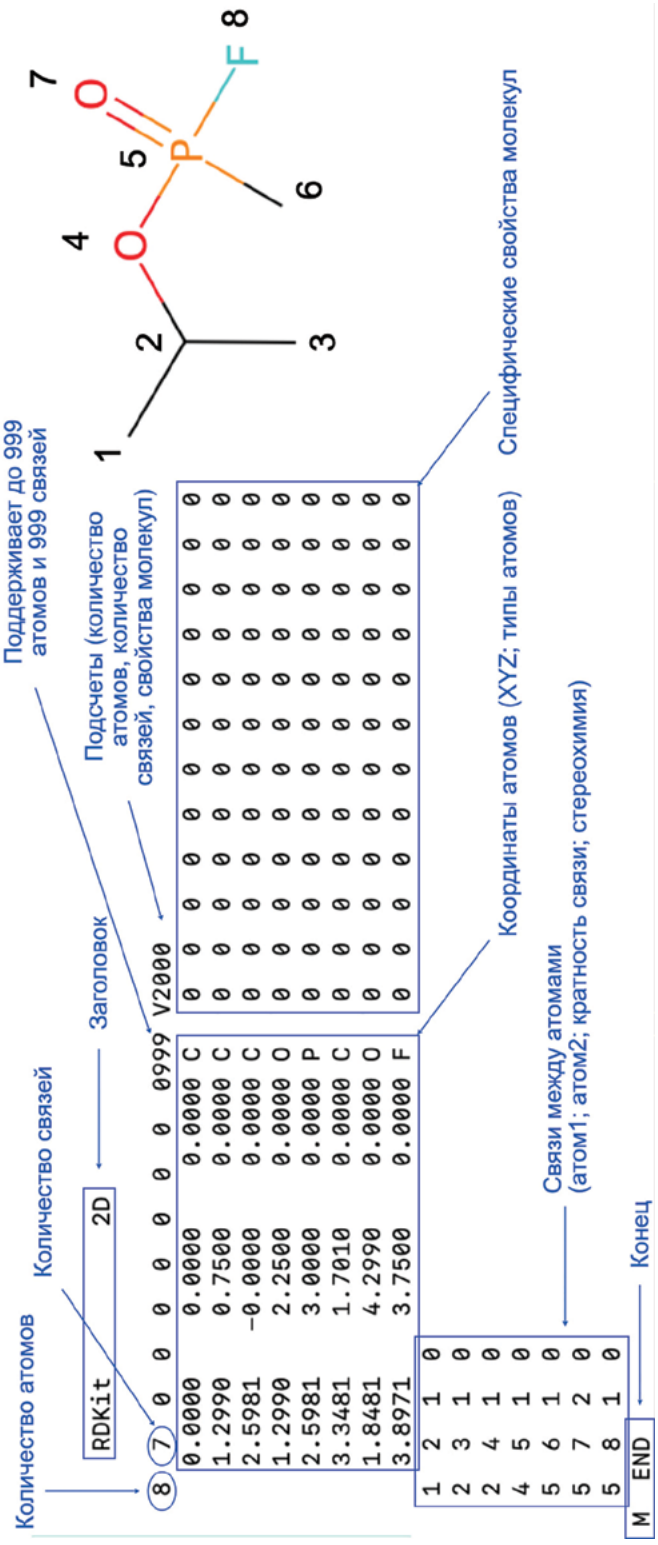


Рис. 17. Структура Зарина в формате MOL

На приведённом выше рисунке видно, что файл MOL состоит из нескольких блоков. В основном блоке содержится информация о координатах атомов, в блоке ниже представлена взаимосвязь атомов между собой и возможная стереохимия (0 – если нет, 1 – если есть). А также блок справа, в котором могут быть отражены дополнительные свойства атомов. Существуют две версии этого формата: V2000 (старая) и V3000 (новая), наиболее популярные программы – визуализаторы химических структур поддерживают оба этих формата.

В случае с форматом MOL2 пользователю доступны возможности более полного описания молекулы, включая частичные заряды атомов и их гибридизацию (рис 18).

В свою очередь формат PDB файлов более массивен, поскольку вынужден содержать информацию о тысячах атомов, которые присутствуют в аминокислотах белковых структур и различных кофакторов, присоединенных к нему вместе с молекулами воды вокруг. Эти аминокислоты кодируются в файле с помощью последовательности из 3 букв, молекулы воды записываются как HOH, а кофакторы и лиганды отображаются в виде специальных символов.

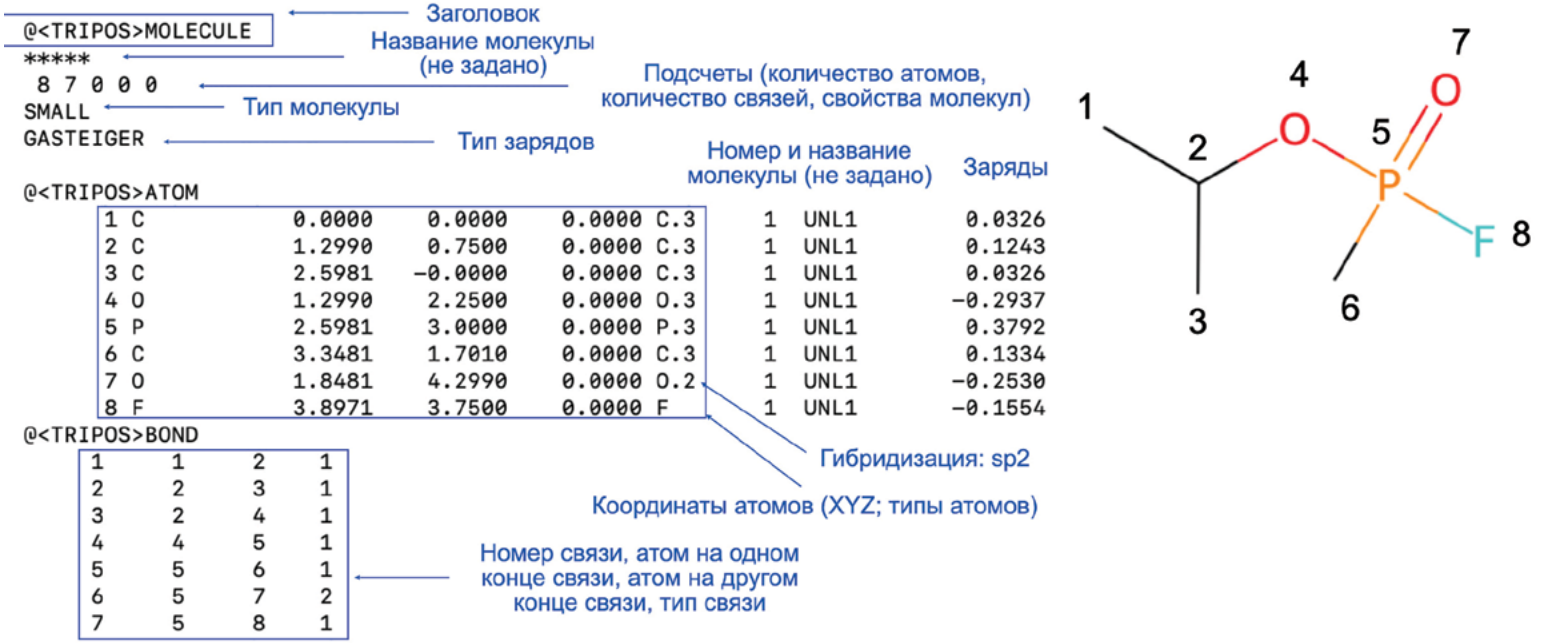


Рис. 18. Структура Зарина в формате MOL2

2.1.4. Переходы между форматами

Поскольку типов цифровых операций, которые можно проводить при работе с молекулами, целое множество – под каждую задачу были разработаны необходимые форматы. Не всегда нужно рассматривать молекулу с точки зрения ее трехмерного представления, а иногда без этого не обойтись. Или не всегда нужно знать о зарядах, располагающихся на атомах, хотя в некоторых исследованиях это может стать критически важным для понимания свойств показателем.

Чтобы не создавать одну и ту же молекулу каждый раз заново в новом формате, были придуманы конвертеры из одного формата в другой. Подобно тому, как различные интернет-ресурсы позволяют переделать документ с расширением .doc в расширение .pdf, химические конвертеры помогают исследователям сократить время и переводить молекулу из одного формата в другой, параллельно проводя необходимые расчёты недостающих показателей при переходе.

Наиболее популярным софтом, который позволяет проводить такие операции, является OpenBabel [31]. Это бесплатная программа, которая позволяет конвертировать более 100 различных химических форматов, включая MOL, MOL2, SDF, PDB, SMILES и многие другие. OpenBabel не просто меняет расширение файла. Программа анализирует структуру молекулы, восстанавливает недостающие данные, такие как координаты атомов или заряды, и обеспечивает корректное отображение всех химических связей, в том числе стереохимии. Она легко интегрируется в рабочие скрипты, поскольку имеет поддержку командной строки. Кроме того, у программы есть GUI (Graphical User Interface – графический интерфейс пользователя), в котором также можно проводить все необходимые операции.

2.2. Базы данных

Одним из главных достижений хемоинформатики являются электронные базы данных. Химикам по всему миру было необходимо обмениваться информацией: искать и скачивать для дальнейшего применения в расчётах цифровые версии молекул в различных форматах, узнавать их физико-химические свойства и без ограничений управлять большим объемом данных. А данных к нашему времени скопилось немало. К примеру, сейчас одной из крупнейших химических баз данных является зарубежная платформа PubChem [32]. Она содержит информацию о 121 миллионе соединений (июнь 2025 года). Если бы мы задались целью сделать ее бумажную версию из листов формата А4 таким образом, что на каждой странице приводилась бы информация для 10 молекул (что довольно плотно), то толщина справочника оказалась бы сопоставимой с высотой Эвереста.

В связи с этим, электронные базы данных стали вынужденной необходимостью для эффективной работы с таким объемом информации. В них реализованы специальные алгоритмы поиска, которые должны эффективно обрабатывать входящие запросы от пользователей и выдавать нужные им сведения за короткое время. В свою очередь, выдаваемая информация должна быть надеж-

ной, поэтому источники сбора всех физико-химических показателей молекул должны быть проверенными и авторитетными.

Чтобы всем этим критериям отвечать, электронные версии баз данных претерпели немало изменений с начала своего существования и совершенствуются до сих пор. Кроме того, как и раньше, они требуют постоянного обновления информации, касающейся новых молекул, появляющихся каждый день, а также их свойств и химических реакций. Фронт таких работ огромен, даже по скромным оценкам, каждый день по всему миру в среднем появляется порядка 10^4 новых молекул. Такое количество не может не подчеркивать важность быстрого реагирования и актуализации этой информации буквально в режиме онлайн. В условиях стремительного роста объема информации базы данных становятся не просто хранилищами информации, но и мощными инструментами для научных исследований. Некоторые из наиболее продвинутых позволяют исследователям не только находить необходимые соединения, но и на основе встроенных в онлайн платформу алгоритмов машинного обучения предсказывать их свойства, что существенно ускоряет процесс разработки новых лекарств и материалов.

Но к чему же такая спешка? Предсказательная способность моделей на таких платформах в том числе позволяет оценивать токсикологический профиль молекул, т.е. насколько та или иная молекула может быть опасна для человека и экосистемы в целом. На сегодняшний день неизвестны токсикологические свойства примерно 99,8% химического континуума. В день происходит около 10 000 открытий новых веществ, в год – 10^6, однако международные химические базы данных, такие как ChEMBL и CEBS, ежегодно пополняются примерно на 6000 и 400 записей о токсичности соответственно [33]. Разница между скоростью синтеза новых молекул и анализом их физико-химических и токсикологических свойств составляет минимум 3 порядка. Поэтому только с помощью экспериментальных исследований сложно добиться системного понимания о возможных химических рисках для живых организмов.

Таким образом, эффективные цифровые базы данных могут стать не только хранилищем знаний, но и инструментом для обеспечения безопасности химических соединений.

2.2.1. Сравнительный обзор баз данных

Как уже упоминалось, базы данных могут содержать разнообразную информацию о структуре химических соединений и их свойствах. Некоторые из этих баз специализируются на токсикологии, предоставляя сведения о потенциальной опасности веществ для здоровья человека и окружающей среды. Некоторые содержат в себе информацию о биологической активности соединений, что сильно помогает при разработке лекарств. Другие базы данных делают упор на физико-химические свойства молекул: растворимость, температура кипения / плавления, молекулярная масса, ADME – свойства (о них мы узнаем позже) и т.д.

В большинстве случаев информация о структурах содержится в форматах SMILES или InChI. Однако также их можно идентифицировать с помощью названия ИЮПАК или номера CAS (Chemical Abstracts Service) – уникального числового идентификатора химических соединений, полимеров и белковых структур [7]. Чаще всего для поиска веществ хемоинформатики пользуются нотацией SMILES, которая поддерживается абсолютным большинством баз данных. В случае с большими белковыми структурами (Protein Data Bank) используют названия биологических мишеней или препаратов, действующих на них.

Разнообразие таких баз дает возможность пользователю выбирать наиболее удобный вариант, начиная от интерфейса, заканчивая типом файлов, которые необходимо скачать для дальнейших исследований. Среди наиболее популярных можно привести в пример PubChem, ChEMBL, ZINC, DrugBank, и другие. В следующих главах мы рассмотрим подробнее некоторые из них. Если же вам интересны цифровые механизмы поиска молекул внутри баз данных, их внутреннее устройство и техническая часть, с помощью которой реализуются все процессы, то рекомендуем к прочтению второй том Т.И. Маджидова и коллег «Введение в хемоинформатику. Химические базы данных» [34].

2.2.2. PubChem

PubChem – это открытая база данных химической информации, управляемая Национальным институтом здравоохранения (National Institutes of Health – NIH) в США [35]. Термин «открытая» подразумевает возможность размещения собственных научных данных на платформе, а также свободный доступ к этим данным для других пользователей. С момента своего запуска в 2004 году PubChem стал важным ресурсом для учёных, студентов и широкой общественности, предоставляя информацию миллионам пользователей по всему миру. В этой базе в основном представлены малые органические молекулы, но также имеется информация о более крупных соединениях – нуклеиновых кислотах, углеводах и пептидах. PubChem собирает информацию о химических структурах их идентификаторах, физико-химических свойствах, биологической активности, патентах, данных о безопасности, токсичности и многом другом. Эти данные поступают из совершенно разных источников. Среди них есть и правительственные агентства, и поставщики химической продукции, и, конечно же, издатели научных журналов. Количество структур в PubChem постоянно увеличивается, на данный момент эта база насчитывает около 121 миллиона уникальных химических структур, а также порядка 250 тысяч биологических мишеней, что является довольно большим объемом.

2.2.3. DrugBank

DrugBank был основан в 2006 году в лаборатории доктора Дэвида Уишарта в Университете Альберты (Эдмонтон, Канада) как проект, направленный на помощь научным деятелям в получении подробной структурированной

информации о лекарствах. В 2011 году он стал частью Инновационного центра метаболомики (TMIC). Поскольку популярность проекта продолжала расти, в 2015 году он был выделен в отдельную компанию OMx Personal Health Analytics Inc [36].

Первоначальный проект DrugBank финансировался Канадским институтом исследований в области здравоохранения и Инновационным центром метаболомики (TMIC), который поддерживает широкий спектр передовых исследований в области метаболомики на национальном уровне. TMIC финансируется компаниями Genome Alberta, Genome British Columbia и Genome Canada – некоммерческой организацией, возглавляющей национальную стратегию геномики Канады при финансовой поддержке федерального правительства.

DrugBank Online представляет собой бесплатную онлайн платформу, содержащую информацию о лекарственных препаратах и их мишенях. Этот ресурс сочетает в себе элементы биоинформатики и хемоинформатики, предоставляя детализированные данные о лекарствах, включая их химические, фармакологические и фармацевтические характеристики, а также исчерпывающую информацию о лекарственных мишенях, таких как аминокислотные последовательности и т.д. Всего в базе содержится более 500 000 известных молекул, которые так или иначе связаны с биологической активностью.

Внимательный читатель сразу увидит разницу между количеством данных в предыдущей базе и этой. Но не стоит сравнивать количество «потенциальных» лекарственных агентов и зарегистрированных утвержденных лекарств. Чтобы молекула стала полноценным лекарством или хотя бы перспективной разработкой, она проходит множество стадий, на которых в процессе может сильно видоизмениться. Одна из основных стадий на этом пути называется «from hit to lead», что в грубом переводе означает «от соединения – хита к соединению – лидеру» [37]. В связи с тем, что в DrugBank содержится большое количество именно лекарственных молекул, он находит широкое применение в фармацевтической промышленности и используется химиками-фармацевтами, фармацевтами, врачами и студентами.

2.2.4. ZINC

У каждой базы данных есть определенное позиционирование и плюсы, которыми их разработчики действительно гордятся. В случае с базой данных ZINC основным преимуществом называют наличие 3D структур всех добавленных в базу соединений с расставленными в них биологически значимыми состояниями протонирования (если таковых несколько, то приведено несколько примеров). Также для одной молекулы учитываются несколько таутомерных и конформационных состояний [38].

Как утверждают разработчики, файлы со структурами соединений сразу подготовлены к докингу в популярных для этого программах. Стоит учесть, что такая тщательная подготовка позволяет исследователям эффективно ис-

пользовать эти данные для анализа взаимодействий с биологическими мишенями и оценки потенциальной активности молекул. Обычно учёным приходится проводить ряд операций с файлом молекулы, для того чтобы подготовить ее к докингу, поэтому расчёт перечисленных показателей является существенным подспорьем для ускорения исследований.

В ZINC представлены коммерчески доступные молекулы со ссылками на каталоги поставщиков, у которых можно купить данное соединение. Сейчас база данных содержит 23 миллиона таких соединений. В современных условиях, когда синтез молекул осуществляется достаточно быстро, это создает удобные возможности, поскольку доступ к таким ресурсам позволяет значительно ускорить процесс поиска и получения необходимых соединений для дальнейших исследований и разработки новых лекарств [38]. Это особенно важно в контексте фармацевтической разработки, где скорость и эффективность могут оказать значительное влияние на общую успешность проекта.

2.2.5. PDB

Protein Data Bank (PDB) – это уникальная база данных с удобным интерфейсом, содержащая в себе информацию о 3D структурах биологических макромолекул (белки, ферменты, рецепторы, РНК, ДНК и т.д.). PDB был оцифрован в 2003 году и с тех пор всегда являлся открытым источником, благодаря которому учёные по всему миру могли использовать и обмениваться информацией о структурных особенностях тех или иных биологических мишенях [39].

Почему важно иметь информацию не только о лекарственных средствах, но и белковых структурах, на которые они действуют? Дело в том, что докинг является процессом, моделирующим возможное присоединение лекарства к биологической мишени в определенном месте. В таком процессе важна каждая деталь:

1) какой участок белка рассматривается как перспективный для связывания с потенциальным лекарственным агентом;
2) пригодна ли мишень в принципе для такой операции;
3) насколько хорошо подготовлена цифровая структура данной мишени;
4) какие у нее биологические функции и где она расположена в организме.

Все эти детали требуют тщательного анализа, не говоря уже о физико-химических свойствах, которые также присущи макромолекулам.

В PDB собрана информация из различных источников по каждой из представленных там мишеней, которые в свою очередь сгруппированы по определенным классам: белковые семейства; организмы, из которых они были выделены и т.д. В базе содержится порядка 230 тысяч структур биологических мишеней, позволяющих проводить цифровые исследования в самых разных областях, начиная от сельскохозяйственного сектора, заканчивая зоологией [39].

Таким образом, целевая аудитория такой базы данных довольно широкая, что не может оставить без внимания столь популярный онлайн сервис.

2.2.6. ChEMBL

ChEMBL, пожалуй, это классика медхимической базы данных. Ее основным козырем считают информацию о биоактивности, извлеченную из статей по медицинской химии. С недавнего времени в ChEMBL добавили данные с клинических и доклинических испытаний, в частности, данные о метаболизме лекарств.

Поскольку процесс разработки терапевтических агентов неприлично дорогой и рискованный (>1 млрд $), наличие информации о причинах отклонения той или иной молекулы от желаемых показателей может послужить для исследователей хорошей подсказкой о том, какие структурные недочеты в соединении стоит устранить, и двигаться в правильном направлении. В том числе, для лекарственно-подобных молекул в ChEMBL рассчитана сходимость с правилами Липински:

1) не более 5 донорно-водородных связей;
2) не более 10 акцепторно-водородных связей;
3) молекулярная масса менее 500 а.е.м.;
4) коэффициент распределения октанол-вода (log P) менее 5.

Стоит помнить, что правила Липински далеко не являются гарантией того, что молекула станет лекарством. Более того, есть много примеров, которые противоречат этим правилам, но всё равно прошли клинические испытания и успешно продаются по всему миру (антибиотик тетрациклин).

На данный момент база данных ChEMBL содержит более 2,1 миллиона уникальных химических соединений и активно используется медицинскими химиками из-за своей направленности на сбор данных о взаимодействии терапевтических агентов с мишенями [40].

2.2.7. Синтелли

Платформа Синтелли является незаменимым помощником для всех исследователей в Российской Федерации, которые занимаются химической информатикой и смежными химическими науками. В условиях, когда санкции влияют на доступ российских учёных к зарубежным химическим базам данных и научной литературе, необходимо иметь собственную большую, постоянно обновляющуюся и надежную базу данных. Это полностью российская разработка, которая не просто является скоплением химической информации. Синтелли – это целая технологическая ИИ платформа в области органической химии, которая позволяет эффективно искать, управлять и анализировать полученную информацию о химических соединениях.

На момент 1 декабря 2025 года в ее арсенале имеются данные о более чем 160 млн молекул, 150 млн научных публикаций, 7 млн химических реакций, 16 млн патентов и около 13 миллиардов молекулярных данных [41]. База данных Синтелли собрана из большого количества источников, включая упомянутую выше PubChem, TOXRIC [42] и др. Дополняется с помощью собственного алгоритма нейросетевых модулей для извлечения информации из научных документов.

Важной особенностью Синтелли является интеграция более 80 моделей на основе нейронных сетей для предсказания физико-химических, токсикологических, биологических и экологических свойств соединений. То есть буквально в браузере, на платформе, можно предсказать свойства целевых молекул.

Кроме этого, в Синтелли встроены дополнительные модули машинного обучения для целого ряда задач. Например, один из модулей позволяет прогнозировать синтез / ретросинтез интересующих вас химических структур, а также рассчитывать его стоимость. Другой модуль позволяет предсказывать спектральные данные ядерного магнитного резонанса (1H, 13C, 15N и 19F) для малых органических молекул. Следующий модуль способен из изображений молекул генерировать SMILES (что оказалось неожиданно очень полезным инструментом согласно обратной реакции от пользователей). Всё это было бы невозможно без должного количества качественных данных. Однако одним из самых главных достоинств такой платформы является возможность создания своего отдельного от общей базы датасета с возможностью его картирования на химическое пространство. Такие датасеты, к примеру, могут иметь фармацевтические компании, которые из огромного количества химических структур выделяют те, которые интересны им с точки зрения дальнейшей разработки. Все датасеты в такой платформе строго конфиденциальны, и даже разработчики Синтелли не могут просмотреть его содержание. Для крупных компаний это критически важный вопрос, поскольку любая разработка строго конфиденциальна, во избежание утечки ценной информации.

Вопросы, связанные с картированием и навигацией в химическом пространстве с помощью инструментов Синтелли будут подробно освещены в заключительных главах учебника, связанных с практическими приложениями химической информатики в различных отраслях. Но уже сейчас можно сказать, что такой инструмент становится всё более востребован на российском рынке, и, например фармацевтическая компания «ИНФАМЕД К» уже разрабатывает новые лекарственные препараты с помощью платформы Синтелли.

2.3. Представление белковых макромолекул

В предыдущих главах мы уже немного затрагивали вопрос о важности представления больших белковых структур в цифровом формате и то, как это влияет на процесс открытия лекарств. В этой главе мы подробнее рассмотрим иерархию организации структуры белков и то, как их оцифровывают учёные для дальнейших исследований. Читателям, которые захотят более глубоко изучить эту область, мы рекомендуем изучить классические труды [43, 44], одним из которых является курс лекций Фильценштейна А.В. и Птицына О.Б. «Физика белка» [44].

Сам по себе белок представляет собой последовательность аминокислот, которые соединены между собой пептидной связью. Раскроем приведённые термины чуть подробнее.

Аминокислоты – это структурные единицы в белках. Они соединены друг с другом в определенные, уникальные для каждого белка последовательности пептидной связью. И это во многом определяет структуру, а также функцию биологической мишени в организме [45]. Сами же аминокислоты представляют из себя органические кислоты, содержащие аминогруппу при альфа углеродном атоме. От него же отходит уникальный для каждой аминокислоты радикал, который помогает ее идентифицировать (рис. 19).

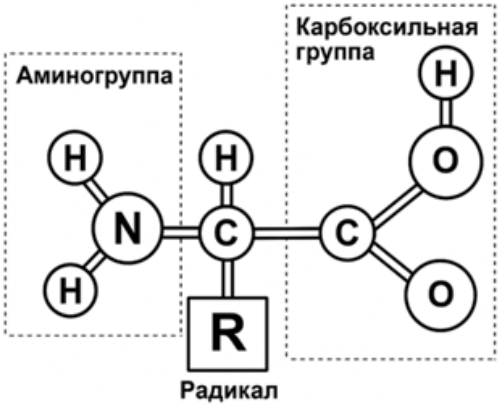


Рис. 19. Общая структура альфа-аминокислот

Всего для построения белков в организме участвуют 20 аминокислот. Стоит помнить, что 9 из них считаются незаменимыми [43]. Этот термин подразумевает, что они не могут быть синтезированы самим организмом, а должны поступать извне (с пищей и т.д.). В таблице ниже приведены все аминокислоты, жирным выделены незаменимые.

Табл. 6. Название основных аминокислот и их аббревиатура [43]

Аминокислота	**Аббревиатура (три буквы / одна буква)**
Глицин	Gly / G
Лейцин	**Leu / L**
Тирозин	Tyr / Y
Серин	Ser / S
Глутаминовая кислота	Glu / E
Глутамин	Gln / Q
Аспарагиновая кислота	Asp / D
Аспарагин	Asn / N
Фенилаланин	**Phe / F**

Окончание таблицы 6

Аминокислота	Аббревиатура (три буквы / одна буква)
Аланин	Ala / A
Лизин	**Lys / K**
Аргинин	Arg / R
Гистидин	**His / H**
Цистеин	Cys / C
Валин	**Val / V**
Пролин	Pro / P
Триптофан	**Trp / W**
Изолейцин	**Ile / I**
Метионин	**Met / M**
Треонин	**Thr / T**

Пептидная связь между аминокислотами образуется благодаря взаимодействию между аминогруппой одной кислоты и карбоксильной группой другой. В процессе этой реакции происходит отщепление воды, а образующийся продукт называют дипептидом [43].

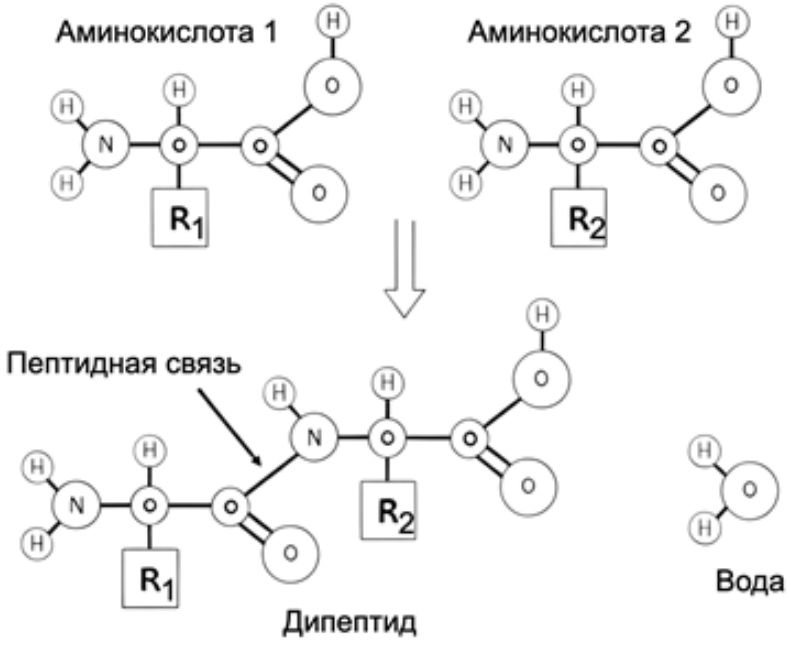


Рис. 20. Реакция образования пептидной связи

Белки являются природными полипептидами, включающими в себя от нескольких десятков до нескольких тысяч аминокислот, соединенных именно такими пептидными связями в составе своей последовательности. Иерархически выделяют 4 вида представления белка: первичная, вторичная, третичная и четвертичная [45]. Об их назначении и пользе такого разбиения порассуждаем в следующих главах.

2.3.1. Первичная структура белка

Первичная структура белка представляет собой конечную последовательность аминокислот, из которых он состоит. С точки зрения биологии, это критически важный уровень организации, поскольку он определяет физико-химические свойства того или иного белка, распределение зарядов в нем и его функции [43]. Каждая присутствующая в нем аминокислота имеет свои уникальные свойства и функциональные группы. Именно они определяют, как белок будет взаимодействовать с другими молекулами, его стабильность, растворимость и активность. Например, полярные аминокислоты, такие как серин и треонин, могут образовывать водородные связи с водой и другими полярными молекулами, что способствует растворимости белка в водной среде. В то же время неполярные аминокислоты, такие как валин и аланин, стремятся избегать воды и могут способствовать формированию гидрофобных взаимодействий внутри белка [43, 45].

Кроме того, последовательность аминокислот напрямую влияет на сворачивание белка в его трехмерную структуру. Правильное сворачивание необходимо для выполнения биологических функций. В свою очередь ошибки в первичной структуре могут приводить к неправильной организации третичной структуры и, как следствие, к утрате функции определенного белка или даже к серьезным патологиям во всем организме, таким как болезни Альцгеймера или Паркинсона [45].

В наших генах хранится информация о том, какая аминокислотная последовательность будет реализована у определенного белка. Не вдаваясь в биологические процессы, просто скажем, что каждая аминокислота кодируется с помощью последовательности из трех нуклеотидов, называемой кодоном. Эти кодоны формируются из нуклеотидов ДНК (аденин – A, тимин – T, гуанин – G и цитозин – C) и определяют, какая именно аминокислота будет включена в полипептидную цепь во время синтеза белка [45].

Каждый кодон соответствует одной из 20 аминокислот или служит стоп-кодоном, сигнализирующим о завершении синтеза. Например, кодон AUG кодирует метионин и также служит стартовым сигналом для начала трансляции.

После того, как синтез белка завершен, на бумаге мы имеем записанную аминокислотную последовательность, что в буквальном смысле и является первичной структурой. К примеру, если мы хотим записать названия аминокислот в формате одной буквы, первичная структура будет принимать вид: -ATKNG - (Аланин – Треонин – Лизин – Аспарагин – Глицин).

Чем же эта информация может быть полезна? Кодирование белковых структур в файлы, которые потом хранятся в базах данных, осуществляется с помощью названий аминокислот из первичной структуры белка. С помощью этих

файлов можно предсказывать трехмерную структуру (Нобелевская премия за «AlphaFold»), изучать эволюционные связи между различными белками, а также выделять ключевые участки для модификации с целью улучшения свойств или разработки новых биомолекул [14].

2.3.2. Вторичная структура белка

Говоря о следующих уровнях организации структуры белка, следует порассуждать: «Почему выделяют четыре уровня организации белка?»

Дело в том, что белковые структуры (рецепторы, ферменты, транспортные белки и т.д.) довольно сложны по пространственному строению. Сотни, а то и тысячи молекул аминокислот соединены друг с другом в единую химическую систему. Каждая из таких молекул имеет свои особенности пространственного расположения атомов и при этом влияет на ближайших соседей, образуя с их же атомами медхимически значимые взаимодействия в разных плоскостях. Всё это нужно учитывать при попытке обобщить для разных белков один и тот же паттерн организации структуры. И обойтись описанием всех пространственных особенностей с помощью единственно верного уровня было бы крайне проблематично. Вторичная структура как раз объясняет строение определенных частей в белках, которые важны при сворачивании в третичную и четвертичную структуры. Таким образом, она относится к локальным пространственным конфигурациям, включая в себя два основных типа: альфа-спирали и бета-слои. Пептидный остов закручивается так, чтобы достичь максимальной стабилизации (снижение свободной энергии) с помощью образования внутримолекулярных водородных связей. В одном белке, как правило, одновременно присутствуют обе структуры, но в разном долевом соотношении. В глобулярных белках преобладает α-спираль, а в фибриллярных – β-структура.

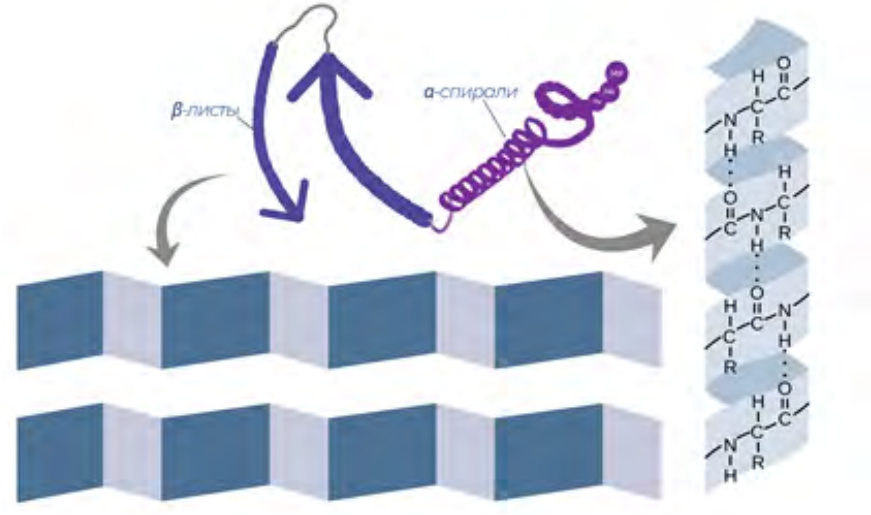


Рис. 21. Вторичная структура белка в форме α-спиралей и β-листов

В α-спирали вторичной структуры полипептидная цепь закручивается в спиральную форму, где каждая аминокислота образует водородную связь с аминокислотой, находящейся через три остатка впереди (1-ая с 4-ой, 4-ая с 7-ой, 7-ая с 10-ой и т.д.). Водородная связь в свою очередь представляет собой связь между кислородом одной группы (акцептором) и водородом другой (донором). На образование такой спирали влияет множество факторов: устойчивость пептидной связи, размер аминокислотного радикала, подвижность связи между центральным атомом углерода и углеродом пептидной группы. При этом боковые радикалы аминокислот максимально удалены друг от друга, а высота одного витка спирали соответствует высоте 3,6 аминокислотных остатков.

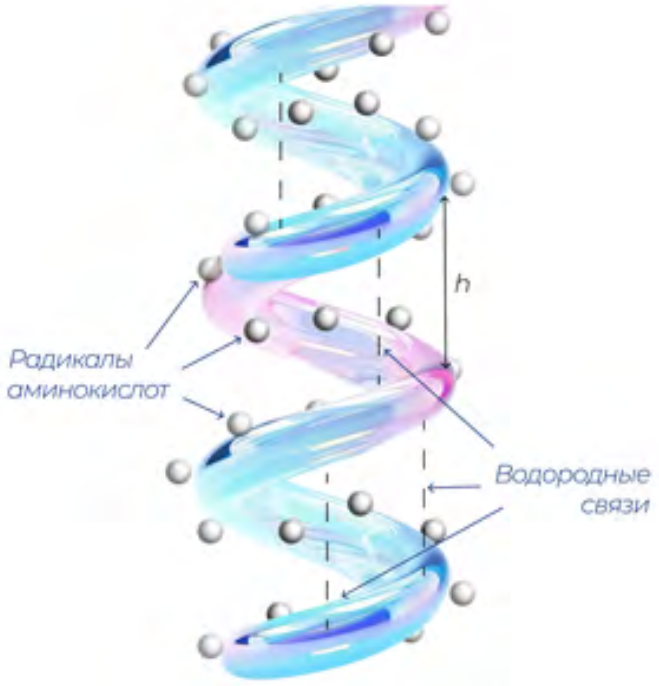


Рис. 22. Схематичное изображение α-спирали

β-слои формируются, когда несколько сегментов полипептидной цепи располагаются параллельно или антипараллельно друг другу, образуя плоские структуры. В таком формате укладки раннее далеко расположенные друг от друга аминокислоты способны взаимодействовать через водородные связи, образуя устойчивую пространственную конфигурацию. Параллельность или антипараллельность определяется направлением белковой цепочки. При параллельной укладке соседние цепи идут в одном направлении. При антипараллельном, соответственно, наоборот. Под направлением, в свою очередь, подразумевают обход последовательности от N-конца (N-концевая аминокислота) к С-концу (С-концевая аминокислота). Таких слоев в структуре белка может содержаться от двух до пяти.

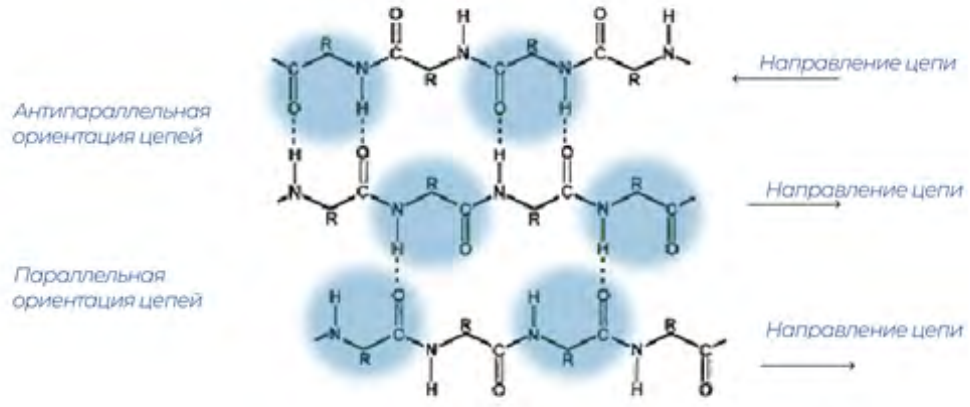


Рис. 23. Параллельная и антипараллельная ориентация аминокислотных цепочек

2.3.3. Третичная структура белка

Строго говоря, все белки в нашем организме можно разделить на фибриллярные и глобулярные [45]. В случае с первыми, основная функция таких белков – структурная, поскольку они входят в состав соединительной ткани (зубы, ногти, волосы, кости и т.д.). Представителей данной категории всего несколько, среди них коллаген, эластин, кератин и другие. Они отличаются сравнительно простым строением и не образуют третичную структуру (фибрилла – нить). Совершенно иначе дело обстоит с глобулярными белками. Для них сложно определить единственно общую функцию, поскольку в эту категорию попадают и рецепторы, передающие сигналы в клетку, и ферменты, катализирующие различные химические реакции в организме, а также транспортные белки и т.д. Однако общий их признак заключается в способности образовывать третичную структуру, принимая которую, они начинают функционировать в полной мере.

Третичная структура – это расположение всех атомов полипептидной цепи в пространстве (рис. 24). Фактически, это 3D модель белка. Стоит подчеркнуть, что в данном случае речь идет об одной полипептидной цепи. Если рассматривать несколько полипептидных цепей, речь уже пойдет о четвертичной структуре (см. внизу).

Почему большинство белков вынуждено существовать в такой форме? Всё дело в гидрофобном эффекте. При образовании третичной структуры в ход идут взаимодействия между боковыми остатками аминокислот (радикалами). Сворачивание пептидной цепочки в глобулу, отдаленно напоминающую шар, происходит благодаря гидрофобным взаимодействиям между неполярными атомами в этих радикалах. Суть таких взаимодействий заключается в стремлении создать термодинамически выгодные для себя условия в полярных растворителях, то есть избегать контактов с водой (или другими полярными молекулами). Таким образом, гидрофобная часть белка находится внутри, а полярная часть, включающая в себя функциональные группы спиртов, аминов и т.д., находится с внешней стороны белковой структуры.

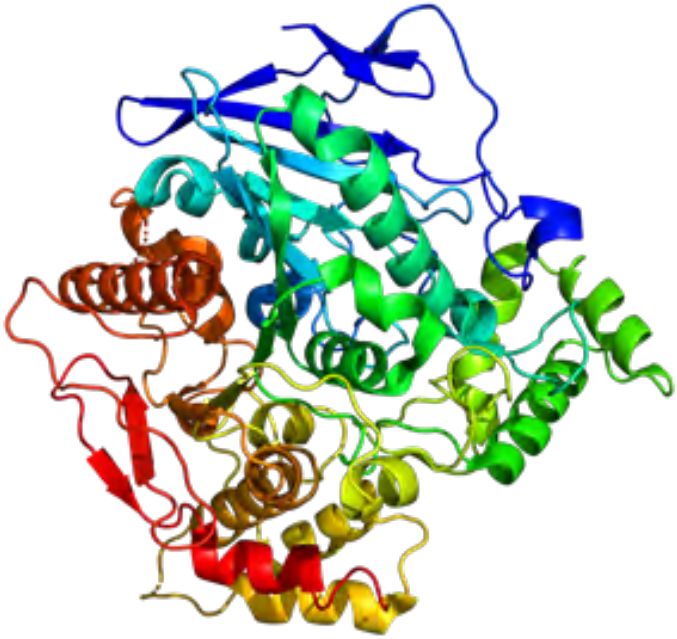


Рис. 24. Пример третичной структуры белка.
Визуализация с помощью программы PyMOL [46]

Отметим, что внутри белка практически нет «свободного места», то есть туда нельзя поместить «ничего лишнего». Однако, места при стыковке доменов (кусков полипептидной цепи) подразумевают наличие так называемых карманов - сайтов связывания. Именно туда проникают молекулы различных лекарств или токсичных веществ, которые действуют на белки, изменяя их конформацию и призывая организм к биологическому ответу [44].

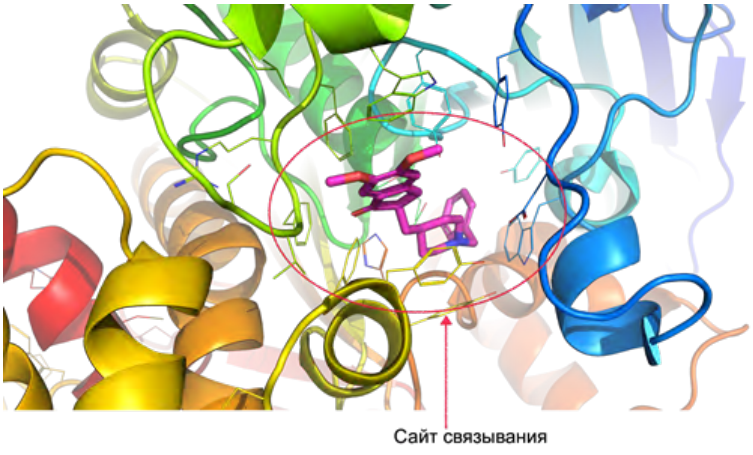


Рис. 25. Сайт связывания в белковой структуре с органической молекулой внутри.
Визуализация с помощью PyMOL [46]

Кроме гидрофобного эффекта в третичной структуре важную роль играют дисульфидные связи или по-другому - дисульфидные мостики. Это ковалентные связи между двумя атомами серы, образующиеся благодаря двум аминокислотным остаткам цистеина. Окисление двух групп тиолей (-SH) в цистеине приводит к появлению ковалентной связи между атомами серы, создавая связь -S-S-. Эти связи помогают стабилизировать трехмерную структуру белка, удерживая его в определенной конфигурации. Дисульфидные мостики могут связывать разные части одной и той же полипептидной цепи. Они способны влиять на активность белка, так как изменения в их числе или расположении могут приводить к изменению формы и функции белка. В некоторых случаях дисульфидные мостики могут открываться и закрываться в ответ на изменения окружающей среды [44,45]. Такие связи особенно распространены в секреторных белках и антителах. Например, инсулин содержит два дисульфидных мостика, которые помогают поддерживать его структуру.

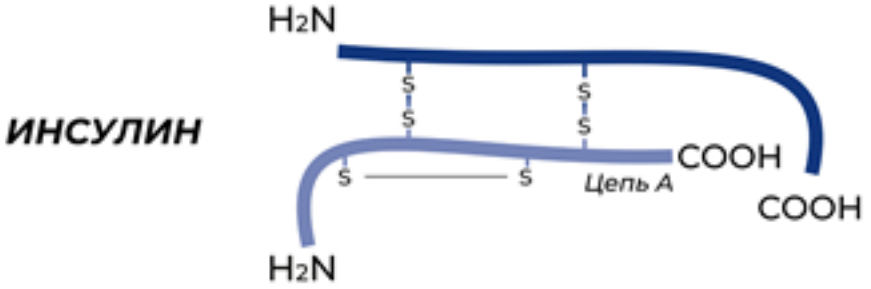


Рис. 26. Схематичные дисульфидные мостики в структуре инсулина

2.3.4. Четвертичная структура белка

В предыдущем разделе мы успели немного проговориться о том, что четвертичная структура белка представляет собой уровень организации, который возникает, когда несколько полипептидных цепей (субъединиц) объединяются в один функциональный комплекс. В среднем таких субъединиц (глобул) может содержаться в четвертичной структуре от 2 до 8, но их число всегда четное. Если структура содержит 2 субъединицы, она называется димером; 4 – тетрамером, 6 – гексамером, 8 – октамером. Эти субъединицы могут иметь одинаковую или различную первичную структуру. Между собой они соединяются с помощью различных взаимодействий, образуя водородные, ионные и гидрофобные связи, а также известные нам дисульфидные мостики. Этот уровень структуры критически важен для понимания того, как белки функционируют в организме [44, 45]. Когда несколько глобул объединяются в четвертичную структуру, они могут выполнять более специфические функции в организме. К примеру, гемоглобин состоит из четырех полипептидных цепей, каждая из которых практически идентична миоглобину. И тот и другой участвуют в переносе кислорода. Однако миоглобин, в силу своей третичной структуры и про-

диктованными ей функциональными особенностями может переносить кислород только по скелетным мышцам. В свою очередь гемоглобин присутствует в эритроцитах и участвует в переносе кислорода от легких к другим органам и тканям по всему организму [43]. На рис. 27 изображена четвертичная структура гемоглобина и третичная структура миоглобина.

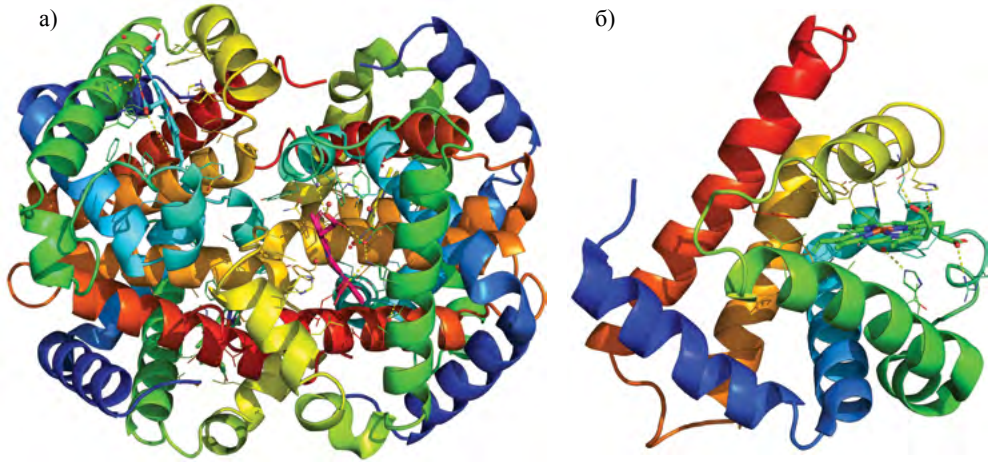


Рис. 27. Четвертичная структура гемоглобина (а) и третичная структура миоглобина (б). Визуализация PyMOL [46]

2.3.5. Инструменты визуализации макромолекул

Безусловно, хемоинформатикам, медицинским химикам и всем смежным специалистам по этим тематикам необходимы знания молекулярной биологии для понимания общей картины происходящих процессов. Однако для химиков-информатиков прикладными исследованиями являются расчёты на ЭВМ. И всё упирается в софт, позволяющий проводить такие расчёты. В случае с разобранными нами структурами белков, этим софтом являются программы для визуализации и обработки макромолекул. Кроме визуализации они часто наделены дополнительными функциональными возможностями, которые позволяют быстро производить несложные подсчеты. Например, расставлять водороды в структуре белков, подсчитывать молекулярную массу биологических мишеней и распределение зарядов в их структуре. Одной из наиболее популярных программ, которая широко используется в академическом кругу, является PyMol. Про другие примеры ПО для визуализации мы расскажем в следующей главе.

PyMOL – это пользовательская система молекулярной визуализации с открытым исходным кодом, поддерживаемая и распространяемая компанией Schrödinger. Примерно четверть всех публикуемых в научной литературе изображений структур белков сделана с помощью этой программы [46]. Ее удобства заключаются в возможности поддерживать различные входные файлы

структур (PDB, SDF, MOL2 и др), а также интеграция с языком Python, что позволяет писать собственные скрипты для автоматизации задач, создания пользовательских функций и расширения возможностей программы. Программа также может использоваться для моделирования взаимодействий между молекулами, что полезно в области дизайна лекарств. На рисунке 28 ниже приведён скриншот рабочего окна PyMOL.

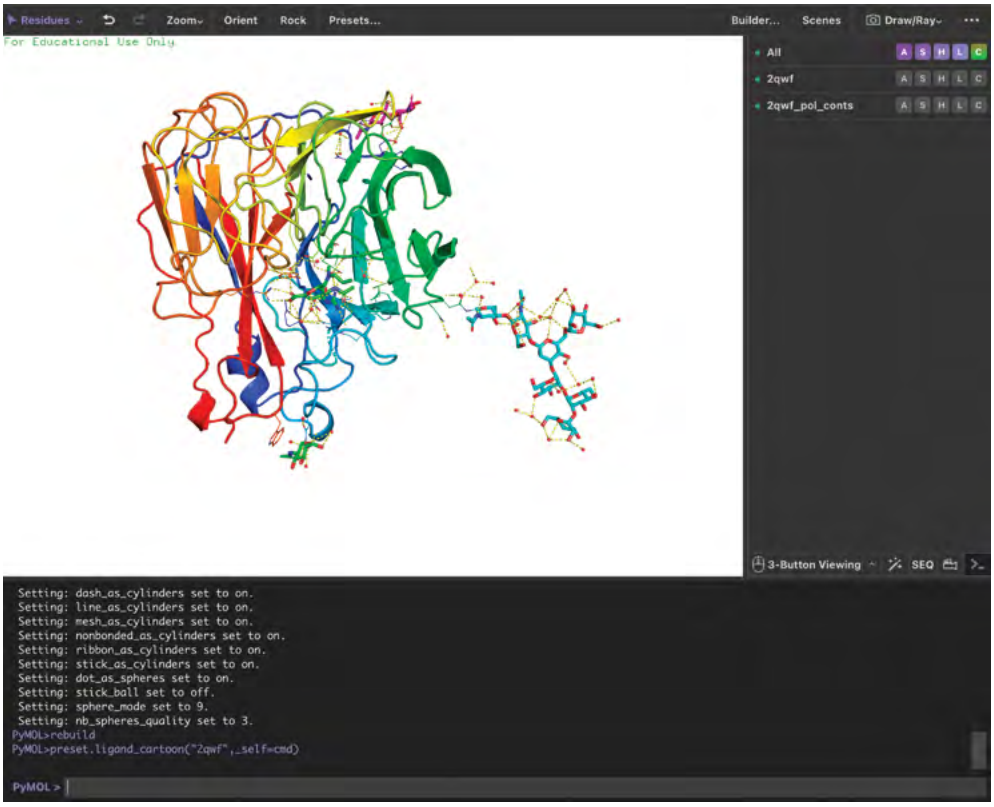


Рис. 28. Интерфейс программы PyMol [46]

3. МЕТОДЫ И ИНСТРУМЕНТЫ ХЕМОИНФОРМАТИКИ

3.1. Машинное обучение (МО)

Прежде всего, в этой главе нам следует разобраться с понятием «машинное обучение». Считается, что впервые его ввел в 1959 году Артур Самуэль. Согласно нему машинное обучение – это способность компьютера учиться, не будучи явным образом запрограммированным [47]. Иначе говоря, основная идея машинного обучения – научиться решать задачу, извлекая закономерности из данных.

Вскоре мы разберем основные виды задач, которые решает машинное обучение. Но прежде стоит упомянуть несколько важных понятий, которые позже будут проиллюстрированы на примерах.

Обучающий набор данных – это коллекция примеров, используемых для обучения модели машинного обучения.

Метки объектов (от англ. *labels*) – это целевые значения или категории, которые присваиваются каждому объекту в обучающем наборе данных и которые модель должна научиться предсказывать.

Обучение модели – это процесс, при котором алгоритм машинного обучения анализирует обучающий набор данных, выявляет в них закономерности и на их основе настраивает свои параметры для решения конкретной задачи.

Теперь, рассмотрим задачи, решаемые с помощью машинного обучения. Их можно разделить на 2 типа: классификация и регрессия (рис. 29)[1].

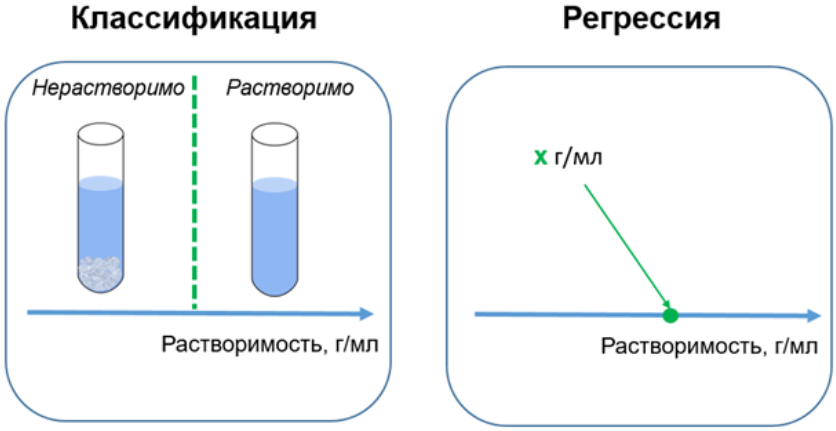


Рис. 29. Классификация и регрессия наглядно

[1] Строго говоря, здесь описываются задачи для обучения с учителем (от англ. Supervised learning). Также существует обучение без учителя (от англ. Unsupervised learning), которое применяется для решения ряда других задач (кластеризация, понижение размерности и др.).

1) Классификация

Классификацией в машинном обучении называется задача, в которой необходимо отнести объекты к одному из заранее заданных классов. Например, мы можем разделить все молекулы на растворимые в воде и нерастворимые. В таком случае, под классификацией молекулы будет пониматься ее отнесение к одному из этих двух классов. Важно, что в действительности молекулы можно разделить на группы по множеству других признаков кроме растворимости: активность к биологической мишени (активна/неактивна), проникновение через гематоэнцефалический барьер (проникает/не проникает), ингибирование фермента CYP3A4 (ингибирует/не ингибирует) и др. Стоит отметить, что во всех упомянутых примерах рассматривалась бинарная классификация, когда мы имеем дело только с 2 классами. На практике, мы нередко сталкиваемся с классификацией веществ более, чем на 2 группы. Например, химические соединения можно разделить на 5 классов опасности, аналогично можно выделить несколько классов растворимости и т.д. [48]. Подобные задачи называются однометочной мультиклассовой классификацией (от анг. single-label multiclass classification) и, к счастью, машинное обучение справляется и с ними [49]. Кроме того, существует более сложный случай мультиклассовой классификации, когда один объект (молекула) может принадлежать одновременно к нескольким классам. К примеру, мы хотим проверить селективность некоторых лекарств по отношению к адренорецепторам. Как известно, существует 2 подтипа α-адренорецепторов (α_1, α_2) и 3 подтипа β-адренорецепторов (β_1, β_2, β_3). В таком случае, каждое лекарство гипотетически может действовать на все 5 подтипов, в то время как наиболее селективные препараты только на 1 из них. Задача классификации в данном случае носит название многометочной (от анг. multi-label). Типы классификаций обобщены на рисунке 30.

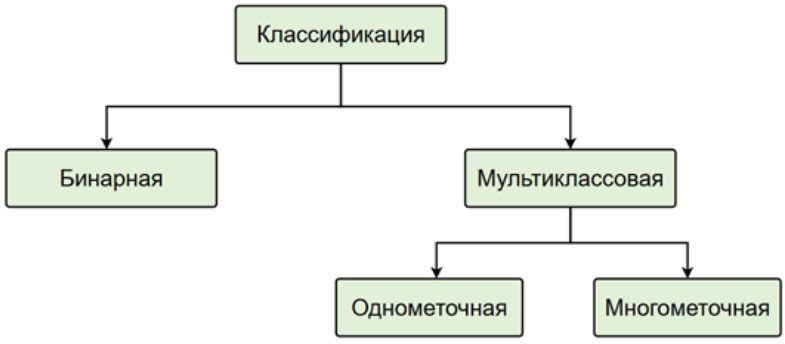


Рис. 30. Типы классификаций

2) Регрессия

Регрессией в машинном обучении называется задача прогнозирования количественной переменной [50]. Если ранее, в случае классификации, мы обсуждали, что можем спрогнозировать растворимо вещество в воде или нет, то в случае регрессии мы можем предсказать значение растворимости, то есть сколько граммов вещества растворяется в единице объема. Это касается и всех

других параметров, измеряемых количественно (коэффициент распределения октанол-вода, константа диссоциации, LD_{50} и т.д.).

Важно понимать, что регрессия и классификация часто идут рука об руку так как для определения метки класса мы вычисляем отношение шансов попадания объекта в разные классы. В задаче бинарной классификации из-за математической простоты, удобства вычисления и интерпретируемости часто применяется логистическая функция, дающая значения от 0 до 1. По порогу 0.5 разделяют шансы отнесения объекта к тому или иному классу. Такой метод называется логистической регрессией, но несмотря на название, используется для классификации.

3.2. Типы данных в машинном обучении

Под данными в машинном обучении понимается множество измерений и наблюдений, которые используются для обучения модели. В случае хемоинформатики их называют дескрипторами. Данные делят на 2 категории:

1. Качественные (категориальные) данные

Эти данные описывают принадлежность объекта к определенной категории. 1) Номинальные данные – это тип категориальных данных, состоящий из категорий/имен, которые не могут быть ранжированы или упорядочены. Например, гидрофобные/гидрофильные молекулы. 2) Порядковые данные – тип категориальных данных, которые могут быть упорядочены. Например, класс опасности вещества.

2. Количественные данные

Это числовые данные, которые можно измерить и упорядочить.

1) Непрерывные данные – могут принимать любые значения в некотором диапазоне[2]. Например, молекулярная масса.

2) Дискретные данные – принимают только отдельные целые числовые значения. Например, количество доноров водородной связи.

3.3. Метки объектов

Как мы помним, под меткой объекта в машинном обучении подразумевается целевое значение или категория объекта. Целевое в том смысле, что именно эту метку мы и планируем прогнозировать после обучения модели. Разберем эту тему более детально в зависимости от решаемой задачи.

1) Метки для классификации

Если мы говорим о модели машинного обучения, которая прогнозирует классы растворимости веществ в воде (растворимо или нерастворимо), то и

[2] Строго говоря, в силу дискретной природы цифровых вычислений на современных компьютерах, вещественные типы тоже имеют конечное число значений, которое определяется форматом внутреннего представления вещественного числа и задаваемой точностью вычислений. Однако количество возможных значений вещественных типов очень велико, поэтому во многих практических задачах конечностью множества на которых представлены данные можно пренебречь (но не всегда!) – подробнее см. [51].

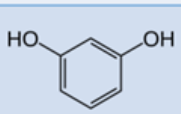
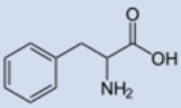
входные данные должны содержать список молекул с известным для каждой из них классом растворимости. Они и выступают метками молекул.

2) Метки для регрессии

По аналогии в задачах регрессии входными данными служит список молекул с известным значением какого, либо значение растворимости, т.е. конкретным вещественным числом.

Различия в метках в зависимости от решаемой задачи проиллюстрированы на рисунке 31.

Классификация

Формула	Класс
	1
	0
	1

Регрессия

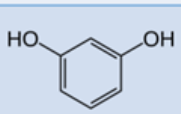
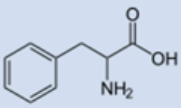
Формула	S, г/100мл
	140
	0.004
	2.8

Рис. 31. Входные данные в случае классификации и регрессии

Важно сказать, что в подобных случаях, как в нашем примере, имея данные о растворимости веществ (вещественное число), мы можем переходить от регрессионной задачи к классификационной. Для этого нам нужно установить порог (например, соединение растворимо, если S > 0,5 г/100 мл) и поменять значения растворимости на классы. Однако ясно, что обратная задача уже не решается, т.е. недостаточно одного знания о классе растворимости, чтобы сказать в каком количестве растворяется вещество.

Следующий вопрос, который может возникнуть, где взять входные данные с известными классами значениями растворимости? Это экспериментальные данные, и они могут быть либо получены самостоятельно или с помощью коллег, либо найдены в соответствующей литературе или в обсуждаемых ранее нами базах данных.

3.4. Методы машинного обучения

Существует множество методов машинного обучения. Их можно разделить на методы классического МО и методы глубокого обучения (различные нейронные сети). Цель данного учебника лишь обеспечить представление о работе машинного обучения, поэтому в этой главе мы рассмотрим только несколько

наиболее простых, но в то же время и сравнительно популярных методов. Для более углубленного изучения предлагаем следующие пособия [52].

3.4.1. Деревья решений

Дерево принятия решений (от англ. decision tree) является простейшим методом машинного обучения. Он используется как для задач классификации, так и регрессии. Дерево решений представляет из себя иерархическую структуру, в которой переход на низлежащие уровни осуществляется по т.н. *решающим правилам.* Иначе говоря, для определения метки объекта мы последовательно задаем вопросы, на которые можно ответить «да» или «нет» [52].

Разберем классический пример, отвлеченный от химии. В таблице 7 приведены данные о группе людей (пол, возраст). В терминологии машинного обучения эта таблица называется тренировочным (или обучающим) набором данных. На его основе мы хотим обучить модель, которая будет прогнозировать наличие у человека зависимости от компьютерных игр на основе его пола и возраста.

Табл. 7. Датасет с метками по игровой зависимости

Пол	Возраст	Играет в компьютер
М	14	1
М	15	0
Ж	10	0
М	9	1
Ж	37	0
М	67	0

В таком случае дерево принятия решений будет выглядеть следующим образом (рис. 32).

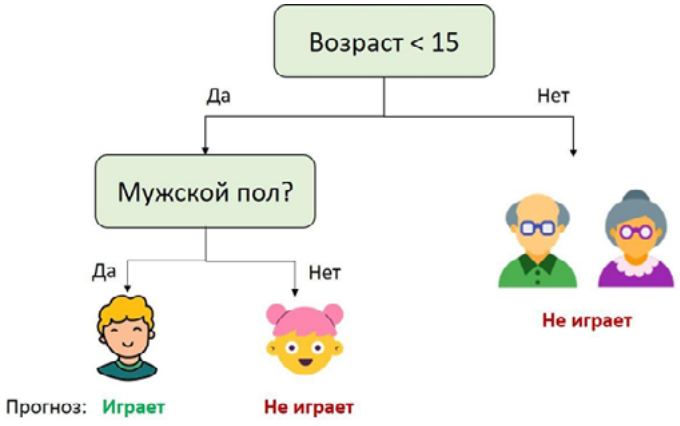


Рис. 32. Дерево принятия решений

3.4.2. Случайный лес

Random Forest или в переводе на русский – «Случайный лес» является важным инструментом в машинном обучении. Его структура представляет из себя совокупность нескольких независимых друг от друга деревьев решений, каждое из которых «голосует» за определенный класс объекта, обучившись на некотором наборе данных (рис. 33). После этого, учитывая каждый голос всех деревьев, принимается окончательное решение о классификации того или иного объекта, поступающего в обученную модель.

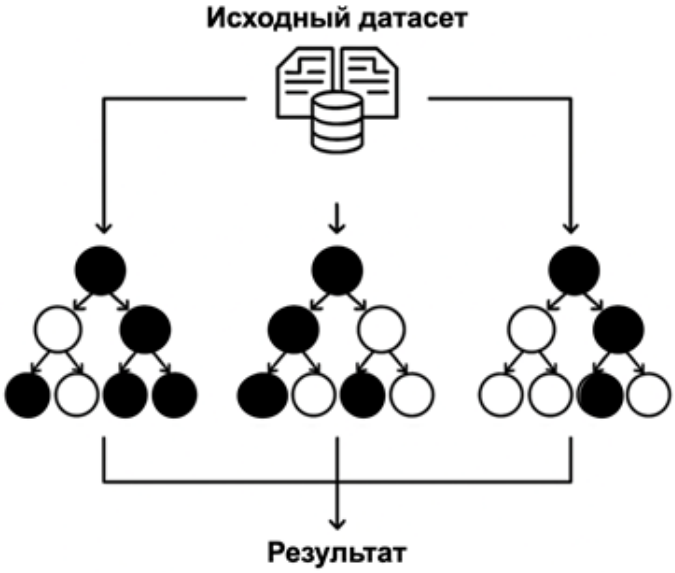


Рис. 33. Схема случайного леса

Представим себе ситуацию. Вы вышли из метро, и вам необходимо добраться до конечной точки маршрута пешком. Ваш телефон, к сожалению, сел и открыть карты не получится. Спросим у прохожих, как вам дойти до назначенного места. Для того, чтобы принять взвешенное решение о том, как вам попасть в нужную точку кратчайшим образом, мы спросим несколько человек. Спросим старожила, который живет здесь половину своей жизни, спросим недавно приехавшего в этот район студента, спросим маму с коляской и ещё пару человек. Каждый из них даст вам свою точку зрения о том, как добраться кратчайшим образом до пункта назначения. Мы выслушаем каждый из вариантов и примем окончательное решение. Такой подход кажется логичным, поскольку есть целая серия вариантов, из которых можно либо выбрать лучший, либо взять лучшее от каждого. Случайный лес работает именно по такому принципу. Однако есть несколько технических деталей, которые стоит обсудить.

Одним из первых и ключевых этапов при построении модели Random Forest, является создание нескольких псевдовыборок из исходного датасета.

Зачем это нужно? Если бы мы отдали один и тот же датасет каждому дереву, наши результаты не отличались бы вовсе. Это как спрашивать несколько раз дорогу у одного и того же человека. Техника, при которой из исходного датасета формируются псевдовыборки, каждая из которых будет обучать свое собственное дерево, называется Bootstrap (Бутстрэп) [53]. Идея такого механизма несложная: у нас есть исходный датасет, из которого мы **случайным образом** «вытаскиваем» какой-то объект, записываем его в выборку и «кладем» обратно (выбор с возвращением). Повторяем такие операции определенное количество раз и получаем сгенерированные случайным образом псевдовыборки, на каждой из которых и будет обучаться соответствующее ей дерево решений (каждому дереву ставится в соответствие своя псевдовыборка). Размер этих выборок обычно соответствует размеру исходного датасета. Поскольку записанные объекты «кладутся» обратно, в одной подвыборке мы можем встретить один и тот же объект несколько раз, а некоторый можем не встретить и вовсе. Эта технология разбиения генеральной совокупности данных позволяет получить разные обучающие выборки из одного и того же исходного датасета, что непосредственно создает разнообразие между деревьями решений.

Следующим важным этапом является механизм разделения узлов в каждом дереве при обучении и предсказании модели. Есть такое понятие, как Bagging (Bootstrap Aggregation) [54], оно является ключевым для понимания различий и отличительной особенности Random Forest от других методов ансамблирования. На самом деле, бэггинг – это процесс, при котором происходит обучение на сгенерированных бутстрэп-выборках. Однако ключевой вопрос заключается в том, как оно происходит? При обучении модели в бэггинге, каждое дерево видит сразу **весь набор признаков**, по которым в нем происходит разделение узлов и конечная классификация объекта. Например, представьте, что 10 врачей диагностируют пациентов. Каждый врач осматривает разные группы пациентов (но некоторые пациенты попадают к нескольким врачам). Доктор использует **все симптомы** для постановки диагноза. А итоговый диагноз определяется по большинству голосов. На этой стадии и проявляется ключевое отличие Random Forest. В нем используется немного другой алгоритм принятия решений внутри каждого дерева. При построении каждого дерева для разделения узла выбирается **случайное подмножество признаков** (например, $\sqrt{n}$ признаков из n для классификации и n/3 для регрессии). То есть, те же 10 врачей, но теперь каждый при осмотре пациента учитывает только **случайный набор симптомов** (например, 3 из 5 возможных). Это заставляет их фокусироваться на разных аспектах, снижая корреляцию между диагнозами. А итоговый диагноз снова определяется голосованием. Добавление случайности в выбор признаков делает деревья в случайном лесу менее скоррелированными и устойчивыми к переобучению.

На этой стадии главное не запутаться. В Random Forest на каждом шаге построения дерева (в каждом узле) случайно выбирается подмножество признаков для поиска лучшего разделения. Это не фиксированный набор признаков на всё дерево, а новый случайный набор для каждого узла. Закрепим, представьте,

что есть группа экспертов (деревьев), и каждый раз, когда эксперт должен принять решение (разделить данные), ему случайным образом показывают только часть всей информации (признаков). Для следующего решения ему снова показывают другую случайную часть информации. Так эксперты принимают решения, опираясь на разные аспекты данных, что делает их мнения менее похожими друг на друга. Отметим также, что Random Forest может применяться как в задачах классификации (голосование), так и в задачах регрессии (усредняя значения каждого классификатора).

Математическое обоснование того, почему каждая из перечисленных стадий работает на улучшение предсказания, можно прочитать в 4 томе Т.И. Маджидова и коллег «Введение в хемоинформатику. Машинное обучение», в которой также обсуждается много других алгоритмов Машинного обучения [55].

3.4.3. Градиентный бустинг

Градиентный бустинг – это также мощный ансамблевый метод, который строит модель итеративно, обучаясь на антиградиенте функции потерь предыдущего блока [56]. То есть строит итоговую модель как последовательность слабых моделей (обычно деревьев решений), каждая из которых последовательно исправляет ошибки предыдущих. В итоге каждая по отдельности слабая модель объединяется в единый мощный механизм, улучшая качество предсказаний.

Без аналогий не обойдемся и в этот раз. Представьте, что Вы играете в гольф. Ваш первый удар – это начальное предсказание, которое далеко от цели. Каждый следующий удар – это попытка исправить ошибку предыдущего, учитывая, насколько далеко мяч от лунки после последнего удара. Вы прицеливаетесь, чтобы уменьшить расстояние, и с каждым ударом приближаетесь к цели. Градиентный бустинг работает так же. Каждая новая модель пытается «поправить» ошибки предыдущих, двигаясь в направлении уменьшения общей ошибки.

3.5. Программное обеспечение для химической информатики – специализированные пакеты для визуализации и обработки данных

Эта глава носит подготовительный характер, в ней мы приведём несколько примеров специализированных пакетов, которые были созданы с целью решить базовые задачи- визуализация химических структур (3.3.1) и обработка химических данных (3.3.2). Это позволит подготовить читателей с минимальным опытом в области к чтению следующих глав и практическому освоению инструментов хемоинформатики. Для тех читателей, у которых уже есть базовый опыт использования пакетов визуализации и обработки данных в химии, эту главу можно пропустить.

3.5.1. Визуализация химических структур

В этом разделе мы продолжим дискуссию о визуализации молекул, начатую в предыдущей главе. Сразу отметим, что программ для визуализации химических структур, как малых, так и больших биологических макромолекул, существует огромное количество. От полностью бесплатных и тех, которые выдают учебную лицензию, до полноценных платных пакетов, часто используемых в больших компаниях. Мы предоставляем читателю право самостоятельного выбора подходящего пакета с помощью поисковых систем или на основе советов коллег. В данном разделе мы хотели бы осветить ряд общих вопросов, в частности – для чего всё-таки нужна визуализация химических структур в условиях бурно развивающихся систем *автоматического* (=с минимальным участием человека) анализа и обработки химических данных?

Дело не только в том, чтобы «поглазеть» на структуру молекулы ради научного интереса. Программы, позволяющие рисовать химические соединения, в первую очередь используются для *интерактивной* работы с компьютерными программами и для генерации изображений в технической литературе, научных и научно-популярных статьях, учебниках и др. С их помощью рисуют необходимые структуры, расписывают реакции и механизмы, а также анализируют тонкие пространственные признаки наподобие углов и расстояний между атомами. То есть фактически, они предоставляют *интерфейс* между компьютерными программами и человеком в плане *осмысления людьми* химической информации.

Одну из таких программ для пространственной визуализации мы уже успели осветить в разделе о структурах биологических мишеней – ею является PyMol [46]. Однако не все программы имеют возможность воспроизводить 3D модели молекул. Да это не всегда и нужно. Из зарубежных программ, стоит отметить популярную программу ChemDraw из пакета программ ChemOffice компании PerkinElmer [57], которой во многих ВУЗах активно пользуются студенты с младших курсов, отлично зарекомендовала себя как инструмент, выполняющий необходимый набор функций в процессе написания химических статей. В ней удобно рисовать различные химические соединения любой сложности и сохранять в нужных форматах. Не все визуализаторы могут похвастаться удобным интерфейсом для создания молекул. ChemDraw имеет интуитивно понятный визуал и не требует много времени для освоения.

Кроме ChemDraw, популярной программой, используемой в научной среде, является Marvin, созданная компанией Chemaxon [58]. Она активно используется для подготовки иллюстраций к научным статьям, моделирования реакций, генерации структурных формул, расчёта свойств молекул и обмена структурными данными между различными программами и базами данных. В программе есть обширная библиотека готовых фрагментов, а также возможность создавать собственные структурные шаблоны, что ускоряет построение типовых структур и реакций. Автоматически выявляются некорректные структуры (неправильная валентность и т.д.), поддерживается работа с изотопами, зарядами,

радикалами и стереохимией. Прямо из Marvin Sketch можно запускать плагины ChemAxon для расчёта физических, химических и спектральных свойств молекул, таких как pKa, logP, масса, поляризуемость и др.

Многие программы-визуализаторы имеют возможность интеграции с MS Office, что позволяет редактировать формулы непосредственно в документе, в котором происходит написание текстов. Высокая конкуренция в этой нише провоцирует компании не останавливаться только на предоставлении графических редакторов структур, но и создавать целые цифровые химические пакеты, в которых каждый блок отвечает за исполнение определенного функционала, который необходим химикам и хемоинформатикам.

К сожалению, многие зарубежные производители программного обеспечения в области химической информатики в недавнее время стали отличаться высокой непредсказуемостью в плане критериев выдачи и продления лицензий российским организациям (даже ВУЗам и академическим организациям). Часть производителей ПО и баз данных вовсе приостановили выдачу лицензий российским организациям из-за международных санкций. Поэтому в настоящее время российским исследователям мы можем посоветовать использовать либо полностью открытое ПО для визуализации химических структур (например, Jmol, Avogadro, или RasMol) либо отечественные разработки (как, например, упоминавшаяся выше платформа Синтелли [41], в которую встроен ряд инструментов 2D визуализации).

3.5.2. Открытая среда для обработки химических данных – RDKit

Говоря о необходимости вести расчёт большого количества показателей, подготавливать химические структуры к внедрению в модели и просто работать с большим количеством химических данных, нельзя не привести в пример RDKit – библиотеку с открытым исходным кодом, предназначенную для работы именно с химическими данными. RDKit встраивается в код на Python, как, например, NumPy или Pandas. Таким образом, Вы можете писать скрипты на популярном языке Python и использовать ссылки на функции RDKit для химических расчётов [59].

История библиотеки RDKit началась в 2000 году, когда Грег Лэндрум и команда Rational Discovery задумались, как сделать анализ химических данных доступнее. В 2006 году проект стал распространяться по лицензии open-source, и с тех пор он превратился в де-факто глобальный стандарт для работы с молекулярными данными. В силу открытого характера проекта, библиотека наполнилась большим количеством имплементированных методов, поэтому рассказать о каждой ее функции и отразить команды, которыми их можно реализовывать, технически невозможно в рамках этой книги. Поэтому мы приводим ссылку на оригинальный электронный ресурс от RDKit, с помощью которого заинтересованный читатель может ознакомиться со всеми возможностями этой библиотеки за сравнительно короткое время при наличии навыков

программирования на Python [60]. Если у Вас никогда ранее не было опыта в программировании, не стоит бояться начинать знакомство с этой библиотекой. Она имеет интуитивно понятный синтаксис команд и не требует глубокого знания Python на этапе первоначального знакомства. Начните с простых действий, как, например, получение изображения молекулы из SMILES:

```
from rdkit import Chem
# Создаём молекулу из SMILES-строки (формат записи структуры)
molecule = Chem.MolFromSmiles("CCO")  # это этанол
# Визуализируем её
Chem.Draw.MolToImage(molecule)
```

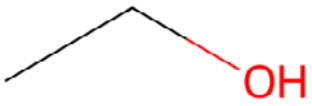


Рис. 34. Молекула этанола, нарисованная с помощью RDKit

Или можно рассчитать молекулярную массу любой интересующей молекулы, приведём в пример Аспирин (ацетилсалициловая кислота):

```
from rdkit import Chem
from rdkit.Chem import Descriptors
# SMILES-строка аспирина (ацетилсалициловая кислота)
aspirin_smiles = "CC(=O)Oc1ccccc1C(=O)O"
# Создаём объект молекулы из SMILES
mol = Chem.MolFromSmiles(aspirin_smiles)
# Рассчитываем молекулярную массу
mass = Descriptors.MolWt(mol)
print(f « Молекулярная масса аспирина: {mass:.2f} г/моль»)
```

В результате такой операции Вы получите молекулярную массу аспирина с двумя знаками после запятой:

Молекулярная масса аспирина: 180.16 г/моль

Ничего сложного, правда? Импортируем необходимые модули и выполняем встроенные в него команды. Начинайте с самых простых действий и постоянно практикуйтесь, со временем вы сможете обрабатывать целые таблицы с тысячами строк SMILES или InChI, преобразовывать их в необходимые форматы и рассчитывать нужные дескрипторы [61]. Для чего это нужно – об этом подробнее расскажем в следующих главах.

4. АНАЛИЗ ДАННЫХ В ХЕМОИНФОРМАТИКЕ

4.1 Прогноз зависимости структура – свойство (QSPR)

4.1.1. Общая концепция

Одной из важнейших задач в хемоинформатике является прогноз химических свойств молекул, которые либо ещё не синтезировали, либо для них отсутствуют экспериментальные данные по измерению необходимых показателей. Такими показателями могут выступать абсолютно любые свойства химических молекул: биологическая активность, токсичность, реакционная способность, растворимость и многие другие.

Благодаря чему можно прогнозировать эти свойства? Основная идея заключается в том, что **похожие молекулы должны обладать похожими свойствами**. На этой стадии возникает ряд концептуальных вопросов: «Что такое «похожесть» молекул?», «Где гарантия, что конкретные изменения в молекуле будут объяснимо влиять на изменения прогнозируемых показателей?», «Какими статистическими методами всё это реализовывать?». На ряд этих вопросов давно дали однозначные ответы, какие-то из них до сих пор являются краеугольным камнем споров в академической среде. В этой книге мы лишь раскрываем базовые концепции и идеи, которые применяются в настоящее время учёными по всему миру. Для более глубокого погружения в тему необходимо прочитать ряд научных статей о специализированных инструментах.

QSAR / QSPR (Quantitative Structure Activity / Property Relationship) – это подход, который благодаря прогностическим моделям позволяет устанавливать количественное соотношение между структурой молекулы и ее активностью / свойством. Под активностью мы будем понимать биологическую активность по отношению к определенной мишени, а под свойством подразумеваем значения той или иной предсказываемой величины, отличной от биологической активности (токсичности, растворимости, летучести и т.д.). Чаще всего в таких методах учёные предсказывают биологическую активность молекулы по отношению к определенной мишени, отсюда и название. В предыдущем разделе мы с вами обсудили методы машинного обучения, именно благодаря этим методам, а также другим статистическим инструментам будут реализовываться модели QSAR [62].

В этом подходе, как и в любом другом, есть ряд ключевых стадий, которые просто необходимо знать и работать с их тонкостями. Рассмотрим их в QSAR.

Одной из важнейших и одновременно самой сложной стадией является сбор и подготовка данных. От их качества напрямую зависит качество построенной модели. В машинном обучении есть нерушимый принцип –

«garbage in, garbage out» (можно перевести как «если ввести мусор, то и на выходе получим мусор»). Это означает, что если в систему вводятся неправильные или некачественные данные, то и результаты, полученные на основе этих данных, будут ненадежными или ошибочными. Поэтому на первой стадии принципиально выбирать данные с «гарантией качества». Почему это такая большая проблема?

Во-первых, не всегда можно найти информацию по необходимым молекулам, которые было бы правильно включить в выборку с целью ее репрезентативности. Если получается собрать полный комплект информации для выбранных структур, то чаще всего эти данные взяты из разных источников. То есть их происхождение неоднородно, а значит экспериментальные протоколы будут разниться, что может существенно изменить обучение модели и ее предсказательную способность. Более того, почти никогда исследователю не удается собрать запланированный объем информации из одной базы данных, что не может не отразиться на результатах. Например, если в выборке недостаточно молекул определенного класса, модель может не научиться правильно распознавать или предсказывать свойства именно этого класса.

Во-вторых, некоторые базы данных часто содержат ошибки в экспериментальных величинах, которые могут различаться с реальными аж на 4 порядка! Часто это связано с банальной путаницей в обозначении приставок (милли / микро) или различиях в экспериментальной методике измерения интересующих показателей. Например, в отдельных лабораториях при измерении IC_{50} могут различаться значения pH, ионной силы, температуры раствора и т.д. Эти и другие параметры могут существенно повлиять на результаты измерения IC_{50}, поэтому важно стандартизировать условия.

В-третьих, подготовка данных часто связана со стандартными, но критически важными операциями, связанными с обработкой массивов данных. Например:

1. Удаление дубликатов: исключение повторяющихся записей для предотвращения переобучения;
2. Очистка от выбросов: удаление или коррекция аномальных значений, которые могут исказить модель;
3. Нормализация: например, преобразование IC_{50} в pIC_{50}, который равен $-\log_{10}(IC_{50})$.

После того как данные были собраны и подготовлены, стоит рассчитать их дескрипторы. Дескрипторами называют любые числовые характеристики молекул, описывающие их физико-химические и другие свойства. Например, молекулярную массу, растворимость, заряды и т.д. Фактически, любое свойство молекулы, которое возможно записать в виде числа, может стать ее дескриптором. Поскольку вычислительной машине необходимо численно преподнести информацию о молекуле, дескрипторы используются для представления молекулярной структуры в виде, пригодном для анализа и построения моделей машинного обучения. О видах таких дескрипторов мы поговорим в следующих разделах.

Также обязательно стоит упомянуть о химической предобработке данных, которая получила на английской языке название data curation. В ходе ее производится анализ содержащихся в обрабатываемой выборке химических структур с целью выявления ошибок ввода и возможных дубликатов, обеспечения возможности расчёта дескрипторов и однозначности получаемых их значений, а также гомогенизации выборки. Например, стандартизованная процедура химической предобработки данных для органических молекул включает следующие этапы:

1. удаление смесей, неорганических и металлоорганических соединений (они обрабатываются отдельно);
2. конвертация структур, удаление солей и выбор состояния ионизации;
3. нормализация специфических хемотипов, резонансных форм и таутомеров;
4. выявление дубликатов;
5. анализ химических структур при помощи интерактивной графики.

Теперь рассчитанные дескрипторы позволяют нам провести количественное соотношение между структурой молекул и их свойствами. У нас есть тренировочный набор данных (соответствие между молекулами и экспериментальными значениями определенной величины) и всё готово к построению модели. На этой стадии необходимо правильно подобрать математический аппарат, который будет в ней реализован. В третьем разделе мы уже обсуждали, какие для этого есть варианты, но стоит помнить, что это далеко не все возможные методы построения предсказательных моделей. Чаще всего, исследователь определяет их применимость на основании линейности данных в тренировочной выборке. Если зависимость между дескрипторами и целевой переменной (например, биологической активностью) линейна, то можно использовать простые линейные модели, такие как множественная линейная регрессия (MLR) или метод частичных наименьших квадратов (PLS) [63]. Они легко интерпретируются и позволяют понять, какие дескрипторы вносят наибольший вклад в прогноз. Если же линейной зависимости не наблюдается, следует попробовать более сложные методы: метод опорных векторов (Support-vector machine – SVM) [64], ансамблевые методы (Random Forest, Gradient Boosting) или нейронные сети [65].

Следующей стадией назовем валидацию модели. По своей сути это процесс оценки ее качества и надежности. Основная цель такого процесса - убедиться в том, что модель корректно работает на данных, которых она не видела во время обучения, но зато их видели Вы. Это необходимо для проверки модели на пере- или недообучение. Наиболее распространенным видом валидации является кросс-валидация. При таком подходе данные делятся на k частей, модель обучается на $(k-1)$ частях и тестируется на оставшейся. Такой процесс повторяется k раз, а полученные результаты усредняются. Преимущества подхода кросс-валидации заключаются в более эффективном использовании данных и независимости от способа их разделения на обучающую и тестовую выборки [66].

Заключительной стадией является применение. Здесь модель выходит за границу нашего с нами знания и выдает ранее не виданные никем результаты.

Стоит осознавать, что это не волшебная таблетка. Одним нажатием кнопки или написанием скрипта не решить все проблемы природы с их разнообразием и квантовыми запутанностями. Однако критическое отношение к результатам и разумное применение новой информации способно расширить или пересмотреть наше представление о текущем положении дел в совершенно разных научных исследованиях химической информатики.

Таким образом, стоит подытожить общий вид QSAR подхода. Это буквальное установление зависимости Y = f(x). Где Y – прогнозируемое свойство (токсичность, активность т.д.), а x – признаки молекул, которые использовались в виде дескрипторов для предсказания целевого свойства. Если же определенной функциональной зависимости между данными не может быть, как в случае применения нейронных сетей или других методов глубокого обучения, будем называть QSAR'ом установление отображения пространства признаков молекул (X_n) на пространство свойств молекул (Y_k), при этом учитывая, что $n \neq k$.

4.1.2. Прогнозируемые свойства

При ответе на вопрос о том, какие свойства молекул являются ключевыми для прогноза их пригодности как лекарственных средств, сразу стоит упомянуть про аббревиатуру ADMET.

ADMET – absorption (абсорбция), distribution (распределение), metabolism (метаболизм), excretion (выведение), toxicity (токсичность) (рис. 35) [67]. Иногда можно встретить альтернативные записи, такие как LADME, где L – liberation (высвобождение) или просто ADME.

Рис. 35. Аббревиатура ADMET

Концептуально, эти термины отражают полный цикл прохождения по организму химического соединения и то, какой след оно может оставить за собой (токсичность). Ежегодное увеличение количества новых структур не привело к ожидаемому росту числа новых лекарств, выпускаемых на рынок. И это, в частности, объясняется плохими фармакокинетическими свойствами [68]. Часто молекулы, проявляющие хорошую биологическую активность, не проходят клинические испытания из-за недостаточно оптимизированной химической структуры, которая может не удовлетворять одному из свойств

ADMET [69]. В свою очередь, большие фармацевтические компании тратят огромное количество ресурсов для тестирования тысяч соединений по каждому из этих показателей. Этот процесс довольно дорогой (клеточные моносло и, химические реагенты, роботы-синтезёры и т.д.), и каждый раз бездумно тестировать химическое пространство, пусть даже суженное до определенного класса веществ, может быть не по силу даже крупному фарм кластеру. Поэтому предварительная оценка потенциальных свойств химических соединений с помощью вычислительных методов стала неотъемлемой частью процесса разработки лекарств [70].

С точки зрения разработки, одним из первых свойств, которое необходимо оценить, является абсорбция в пищеварительном тракте (не путать с адсорбцией (рис. 36)), поскольку степень всасывания препарата через кишечник определяет возможность его перорального применения. Во избежание путаницы, абсорбция – это процесс поглощения сорбата (торта) **всем объемом** сорбента (человек на рисунке кушает торт = концентрация торта **внутри** человека увеличивается), а адсорбция – это процесс увеличения концентрации вещества **на поверхности** раздела фаз (человеку прилетает торт в лицо = концентрация торта на поверхности лица увеличивается) [71].

Рис. 36. Абсорбция/Адсорбция

Два основных фактора, влияющих на всасывание в кишечнике, это растворимость препарата в жидкостях ЖКТ и проницаемость препарата через стенку кишечника. Однако не всё так просто, поскольку растворимость и проницаемость требуют противоположных по логике свойств молекулы. Чтобы молекула была растворимой, например, в желудочном соке, необходимо наличие в ней функциональных групп, которые способны образовывать водородные связи в полярных растворителях, таких как вода или соляная кислота, т.е. молекула должна быть достаточно гидрофильна [72]. В то же время, проницаемость определяется способностью молекулы проходить через липидный бислой мембраны в клетки желудочной стенки (энтероциты). В этом случае соединение, наоборот, должно быть достаточно гидрофобным, чтобы иметь возможность при контакте с мем-

браной с помощью пассивной диффузии проникнуть внутрь клетки [73]. В противном случае, придется рассчитывать на специализированные белки на поверхности мембран для переноса вещества внутрь (активный транспорт).

Для анализа потенциальной способности молекулы перейти из водного раствора в мембрану клетки был введен десятичный логарифм коэффициента распределения октанол - вода «Log P». Сам коэффициент «P» представляет собой отношение концентрации растворенного органического соединения в октаноле к концентрации в воде [73].

$$lg(P_{\text{окт/вода}}) = lg(\frac{[\text{вещество}]_{\text{окт}}}{[\text{вещество}]_{\text{вода}}})$$

Такой параметр по праву может называться дескриптором молекулы, поскольку количественно отражает способность соединения растворяться в полярной и неполярной среде одновременно. Однако насколько он точен? На этой стадии опять возникает проблема качества экспериментальных данных в различных базах. Да, даже в случае с таким простым свойством, как растворимость, хватает неопределенностей [74]. Например, растворимость диклофенака в воде в литературных данных до сих пор может варьироваться до двух порядков [74]. Это обусловлено как структурными особенностями вещества, так и методиками, по которым эту растворимость измеряют. В наше время известно, что при прогнозе растворимости молекулы в организме необходимо также учитывать ряд других факторов (молекулярная масса, площадь гидрофобной поверхности молекулы, количество водородных связей и пр.), а не полагаться только лишь на log P [75].

В случае с распределением лекарственного препарата по организму прибегают к использованию большого количества дескрипторов, поскольку сам процесс многостадийный и подразумевает целый каскад взаимодействий между препаратом и различными биологическими системами. Эти дескрипторы включают в себя такие параметры, как липофильность (Log P), молекулярная масса (MW), связывание с белками плазмы, pH – кислотность среды, pKa – показатель кислотности веществ, объем распределения (Vd) и некоторые другие. Каждый из этих дескрипторов дает информацию о конкретных аспектах поведения препарата в организме, таких как его способность проходить через клеточные мембраны, накапливаться в тканях или связываться с белками плазмы.

В случае с log P мы уже обсудили концептуальный смысл и стороннее применение. Что же касается молекулярной массы, здесь необходимо следить, чтобы молекула не была слишком большой (MW <500), в противном случае ей будет сложно проникать через мембраны клеток или ГЭБ (гематоэнцефалический барьер). Также большие молекулы часто имеют более сложные пути метаболизма, что с большей вероятностью может привести к токсичным метаболитам.

В свою очередь, расчёт pKa дает возможность оценить степень ионизации вещества в различных средах организма [67,69]. Вообще Ka – это константа кислотной диссоциации, которая характеризует кислотность или основность

химического соединения. Ее расчёт происходит по уравнению реакции диссоциации кислоты в растворе на ионы водорода и кислотные остатки. В случае с одноосновными кислотами уравнение и расчёт константы выглядит следующим образом:

$$HA \Leftrightarrow A^- + H^+$$

$$Ka = \frac{[H^+][A^-]}{[HA]}$$

Эта формула показывает отношение произведения концентраций ионов водорода и кислотного остатка к концентрации не диссоциированной в растворе кислоты. Если прологарифмировать обе части по основанию 10 и немного преобразовать выражение, получим:

$$pH = pKa + lg(\frac{[A^-]}{[HA]}),$$

где $pH = -lg([H^+])$, а $pKa = -lg([Ka])$

Данное уравнение называется уравнением Гендерсона – Хассельбаха. Оно позволяет связать кислотность среды (pH) со способностью химического соединения диссоциировать в ней [67,69].

Почему же этот показатель важно учитывать при попытке оценить распределения препарата по организму? Дело в том, что ионизированная форма вещества хорошо растворяется в полярных растворителях, но при этом имеет плохую способность к всасыванию через мембрану. Впоследствии это приводит к неполному попаданию вещества в постоянный кровоток и его распределению по организму [69]. И это лишь малая часть тех дескрипторов, которые используются при моделировании ADMET свойств.

4.1.3. Токсичность как сложный многоплановый феномен

Одним из наиболее сложных для предсказания свойств химических соединений среди всех прочих является токсичность. Это связано с тем, что каскадные механизмы, которые приводят к токсическому эффекту, часто являются многоступенчатыми [76]. Более того, определение токсичности химических веществ важно для выявления их вредного воздействия не только на человека, но также на животных и окружающую среду в целом. В связи с этой многогранностью, сложно дать однозначное *формальное* определение термину токсичности отражающее все возможные детали этого феномена. В XVI веке знаменитый швейцарский врач и алхимик Парацельс сказал: «Всё яд и всё лекарство; то и другое определяет доза.» По сей день мы можем наблюдать явное соответствие этих слов с примерами на лекарственных препаратах и не только.

Например, аспирин, который широко используется как нестероидное противовоспалительное средство, может вызывать серьезные побочные эффекты при превышении терапевтической дозировки (кровотечения, аспириновая

астма) [77]. С другой стороны, многие природные соединения, такие как яды некоторых растений и грибов, в малых дозах могут оказывать терапевтическое действие [78].

Для конкретного примера рассмотрим ботулотоксин – один из самых сильных ядов, известных науке, вырабатываемый бактериями *Clostridium botulinum*. Этот токсин нарушает работу нейромышечных синапсов, что приводит к потере способности мышц сокращаться. В том числе, перестают сокращаться и дыхательные мышцы, из-за чего наступает недостаток кислорода в организме и неспособность самостоятельно дышать. Летальная доза для человека очень мала, она составляет всего около 0,1 – 1 микрограмма при попадании через пищеварительную систему или инъекцию, т.е. для истребления всего населения планеты требуется всего несколько килограммов яда [79]. Но, несмотря на все эти страшилки, отважные девушки (и не только) нашли безопасную для косметических процедур дозировку этого токсина и разглаживают свои морщины с помощью известного всем препарата – ботокс. Такие примеры не могут не подтверждать слова Парацельса и лишний раз подчеркивают сложность термина токсичности.

Так что же такое эта токсичность? Как правило, у многих сложных терминов есть *кажущиеся простыми* определения. Поэтому можно считать, что токсичность – это мера любого нежелательного или неблагоприятного воздействия химических веществ [80]. Однако при таком определении, остается вопрос – а как теперь формально определить эту меру (меры)? Поэтому, в свою очередь, виды токсических эффектов от этого воздействия, которые проявляют соединения, называются конечными точками токсичности (toxicity endpoints – токсикологические эндпоинты). Например, если вещество способно вызвать рак, его конечной точкой токсичности является канцерогенность [81].

Итак, существуют десятки токсикологических эндпоинтов, что подтверждает сложность явления. По этой причине следует разобрать классификации токсических эффектов. Рассмотрим две наиболее популярные системы [82].

I. Системные токсические эффекты.

1) Острая токсичность

Исследования длятся, как правило, не больше недели, а вещество вводится однократно. Обычно измеряется смертность. Наиболее популярным эндпоинтом является LD_{50} (Lethal Dose 50), т.е. доза вещества, вызывающая гибель половины особей из испытуемой группы животных. Важно, что прогностические модели строятся под конкретное животное и конкретный путь введения (например, мыши, *per os*). Также к этой категории относят тест Дрейза (раздражение глаз), раздражение кожи и др.

2) Хроническая токсичность

Продолжительность исследований измеряется месяцами и годами, а длина периода зависит от животной модели. Для грызунов, например, время исследований достигает от 1 до 2 лет. Измеряются изменения веса всего организма или отдельных органов, проводится патогистология и т.д. Эндпоинты: NOAEL, LOAEL и ряд других [82].

3) Канцерогенность

Аномальный рост и дифференцировка клеток, способных привести к раку.

4) Эмбриотоксичность

Негативное воздействие на развивающийся эмбрион или плод.

5) Генотоксичность

Повреждение ДНК и изменение генной экспрессии.

II. Токсические эффекты, специфичные для органа/системы органов.

1) Кардиотоксичность (сердце и сердечно-сосудистая система)
2) Дермальная токсичность (кожа)
3) Офтальмотоксичность (глаза)
4) Гепатотоксичность (печень, желчные протоки, желчный пузырь)
5) Нефротоксичность (почки)
6) Нейротоксичность (центральная и периферическая нервная система)
7) Репродуктивная токсичность (репродуктивная система)
8) Респираторная токсичность (верхние и нижние дыхательные пути).

Многогранность термина токсичность настолько велика, что он является центральным элементом целой научной дисциплины – токсикологии. В свою очередь, на стыке токсикологии и хемоинформатики сформировалась современная наука – вычислительная токсикология [83]. Агентство по охране окружающей среды США (US EPA) определяет ее как «наука о применении математических и компьютерных моделей для прогнозирования неблагоприятных последствий и лучшего понимания отдельных или множественных механизмов, посредством которых данное химическое вещество причиняет вред» [84].

Вычислительная токсикология отличается от традиционной по многим аспектам. К примеру, долгое время тесты на выявление токсических эффектов проводились на животных, затем значимым стали методы *in vitro,* в т.ч. благодаря возможностям высокопроизводительного скрининга. Так, широко известна инициатива Tox21 (Toxicology in the 21st Century) – федеральная программа США (EPA, FDA, NIEHS, NCATS) по высокопроизводительному скринингу [85].

Сейчас же есть возможность сократить дорогостоящие эксперименты с помощью применения вычислительных моделей и, в том числе, частично закрыть вопрос о толерантности использования животных в целях таких тестов [86].

Однако, возможно, самым важным из всех аспектов является масштаб. Масштаб в количестве изучаемых химических веществ, масштаб в широте охвата конечных точек, масштаб в изучаемых уровнях биологической организации, диапазоне условий воздействия, а также в охвате стадий жизни, полов и видов. Каким бы быстрым ни был HTS (High Throughput Screening – высокопроизводительный экспериментальный скрининг), его скорость не сможет сравниться с вычислительной силой современных и будущих ЭВМ.

Наиболее значимыми подходами и инструментами в вычислительной токсикологии являются QSAR; физиологически обоснованное фармакокинетическое моделирование; Read across (прогноз свойств на основе похожего соеди-

нения с известной токсичностью); концепция структурных «алертов» (от англ. *structural alerts* – структурные фрагменты, обуславливающие высокую вероятность проявления соединением нежелательного эффекта) [87].

Некоторыми из основных областей применения вычислительной токсикологии являются

1. определение опасности и приоритетности рисков для химических веществ,
2. установление механизма действия,
3. проверка безопасности пищевых добавок и веществ, контактирующих с пищевыми продуктами,
4. оценка степени вариабельности ответных реакций в популяции людей,
5. оценка токсичности молекул-кандидатов и их метаболитов в ходе разработки лекарств.

Говоря о значимости вычислительной токсикологии, стоит упомянуть, что ее инструменты уже внедрены в практику федеральных агентств ряда иностранных государств: США (EPA, Centers for Disease Control, Food and Drug Administration, National Institutes of Health, Agency for Toxic Substances and Disease Registry), Европа (European Chemicals Agency, Institute for Health and Consumer Protection), Канада (Health Canada, National Centre for Occupational Health and Safety Information), Япония (National Institute of Health Sciences of Japan) [88]. Кроме того, в 2007 году в ЕС был принять регламент REACH (Registration, Evaluation, Authorisation and Restriction of Chemicals), регулирующий безопасность химических веществ [89]. Этот документ стимулирует применение методов *in silico* для оценки профиля безопасности соединений.

Таким образом, в настоящее время наблюдается высокий спрос на развитие методов вычислительной токсикологии, подтвержденный интересом со стороны регуляторов.

4.1.4. Молекулярные дескрипторы

В контексте построения количественных моделей типа «структура–свойство» (QSPR) или «структура-активность» (QSAR), функциональная или *алгоритмическая* зависимость y=f(x) отражает связь между молекулярными признаками (x) и целевым свойством (y). Здесь дескрипторы (x) выступают численными репрезентациями молекулярной структуры, транслирующими её ключевые физико-химические, топологические или электронные характеристики в формат, пригодный для машинного анализа. Часто, при прогнозировании свойств молекул, опираются именно на их **структурные характеристики** и ключевые особенности в них, отвечающие определенным эффектам, в том числе и токсикологическим. Именно поэтому мы рассмотрим такое понятие, как fingerprints, или «отпечатки пальцев» [90]. Название этого термина может слегка запутать читателя, никогда не слышавшего о таком подходе. По своей сути, он отражает определенные структурные фрагменты в молекуле (бензольное кольцо, гидроксильная группа, карбоксильная груп-

па и т.д.), наличие или отсутствие которых мы можем наблюдать. Конечный результат обработки молекулы с помощью «фингерпринта» представляет из себя битовую строку, в которую при наличии определенного фрагмента ставится 1, а при его отсутствии 0 (рис. 37).

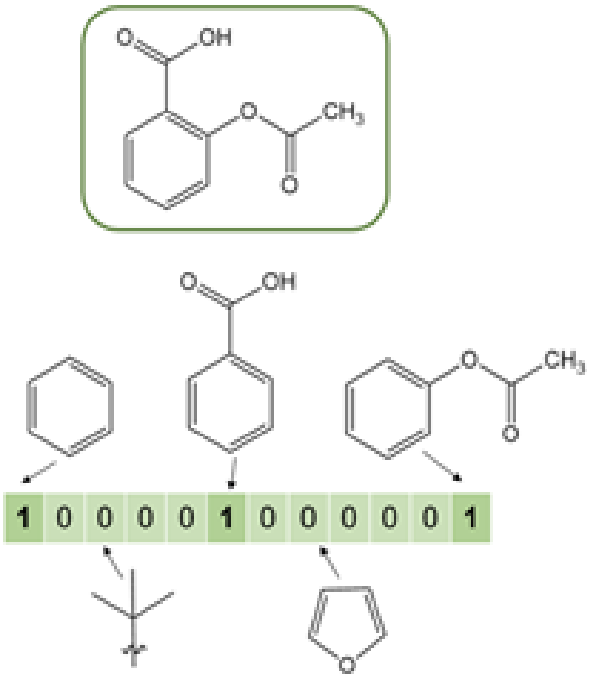


Рис. 37. Пример битовой строки для ацетилсалициловой кислоты

Так работают фингерпринты, которые основаны на *фрагментарной* разбивке молекул. Среди таких большой популярностью пользуются MACCS Keys [85], состоящие из 166 или 960 бит. Каждый бит в строке предопределен структурным фрагментом, на основании которых каждая молекула может быть проанализирована и получить свой булевый массив (битовую строку). Например, 14-ый бит соответствует фрагменту дисульфидной связи, а 125-ый бит соответствует наличию ароматических колец в количестве большем единицы. К сожалению, полностью достоверно неизвестно, какой бит соответствует какому фрагменту, потому что разработаны такие фингерпринты коммерческими компаниями. В случае с MACCS Keys, которые разработала компания MDL, права принадлежат BIOVIA. Версия со 166 бит имплементирована в RDkit, а вот более расширенный формат в 960 бит доступен только в коммерческих пакетах. Расширенная версия может учитывать ряд стереохимических особенностей, сопряженные двойные связи и другие, более сложные фрагменты [91].

Кроме structure-based (структурных) фингерпринтов всё чаще используют path-based (путевые) фингерпринты [92]. Фактически, это фингерпринты, основанные на динамическом анализе путей в графе молекулы. Например, берется атом, и анализируются пути разной длины от этого атома до других. При длине в 1 анализируются непосредственные соседи рассматриваемого атома, при большей длине анализируются не только ближайшие соседи, но и более удаленные атомы, соединенные с исходным атомом через цепочки связей различной длины. Таким образом, для каждого атома в молекуле строятся всевозможные линейные фрагменты – пути, которые могут включать, например, от 1

до 7 связей. Каждый такой путь кодируется определённым образом (например, с помощью хэш-функции) и отражается в битовой строке фингерпринта. То есть, если данный путь встречается в молекуле, соответствующий бит устанавливается в 1.

Классическим примером path-based фингерпринтов являются Daylight fingerprints [93], которые широко используются в задачах поиска структурных аналогов или в ранее упомянутом нами подструктурном поиске. В этих фингерпринтах для каждого атома последовательно перебираются все возможные пути до заданной максимальной длины, а затем полученные фрагменты (например, последовательности атомов и связей) преобразуются в числовые идентификаторы и хэшируются в битовую строку фиксированной длины. Такой подход позволяет эффективно учитывать разнообразие линейных и разветвленных структурных мотивов в молекуле, но при этом неизбежны так называемые битовые коллизии – ситуации, когда разные фрагменты попадают в одну и ту же позицию битовой строки. Чтобы снизить вероятность коллизий, увеличивают длину фингерпринта (например, до 1024 или 2048 бит).

Path-based фингерпринты реализованы во многих современных инструментах, таких как RDKit (RDKFingerprint) и CDK (standard fingerprint) [94], где можно задавать максимальную длину пути и размер битовой строки. Например, в RDKit стандартная длина пути составляет до 5–7 связей, а размер фингерпринта – 2048 бит. Это позволяет гибко настраивать чувствительность метода к деталям в молекулярных структурах.

В отличие от structure-based фингерпринтов, где каждый бит жёстко соответствует определенному фрагменту, path-based фингерпринты строятся динамически для каждой молекулы, что позволяет учитывать уникальные особенности структуры и получать более богатое описание для задач сравнения молекул, поиска аналогов, кластеризации и построения QSAR/QSPR-моделей.

Рассуждая далее на тему фингерпринтов, нельзя не упомянуть о так называемых circular (круговых) fingerprints. Поскольку последнее время они набирают всё большую популярность, разберем механизм их генерации. Они создаются благодаря итеративному описанию окружения каждого атома на основе информации о его соседях **на определенном радиусе**. Ярким примером таких фингерпринтов являются ECFP (Extended – Connectivity Fingerprints) [95]. На первой стадии этого метода каждому атому присваивается идентификатор (уникальный числовой код), который основан на атомном номере, его массе и заряде; числе ближайших соседей, не являющихся водородами; числе присоединенных непосредственно к этому атому водородов; общем порядке связей без учета водородов (валентность минус число связанных водородов) и ещё одном параметре, который проверяет наличие принадлежности данного атома по крайне мере к одному циклическому фрагменту (кольцу). Рассмотрим пример. Допустим, у нас есть молекула глицина (рис 38). Генерируем первичный идентификатор для атома 3.

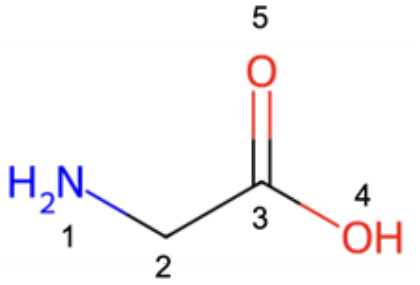


Рис. 38. Нумерация атомов в молекуле глицина

a) Атомный номер = 6
b) Атомная масса = 12
c) Атомный заряд = 0
d) Количество не водородных соседей = 3
e) Число присоединенных водородов = 0
f) Валентность минус число связанных водородов = 4 - 0 = 4
g) Принадлежность к кольцу = 0 (в противном случае – 1)

Таким образом мы имеем уникальный идентификатор атома 3, identifier = (6, 12, 0, 3, 0, 4, 0). Такая операция повторяется для каждого атома и все идентификаторы впоследствии хешируются (например, в Python с помощью команды hash()).

После того, как каждый атом получил свой уникальный идентификатор, информация о нем обновляется с учетом его соседей, которые окружают рассматриваемый атом. Это значит, что после каждой такой итерации идентификатор атомов будет меняться до тех пор, пока не будет пройден заданный пользователем радиус. Обычно он составляет 4–6 атомарных соседей. Например, в ECFP4 реализована версия с расчётом радиуса равного 4, а в ECFP6 равного 6 соответственно. Обязательно ли больший радиус гарантирует лучшие результаты? Ответ – однозначно нет. Важно найти баланс между надежностью и эффективностью. Чем длиннее отпечаток, тем меньше риск совпадений (коллизий), но при этом данные становятся более разреженными. Например, в 1024 битах обычно только от 3 до 30 единиц, а остальные – нули. Удвоение размера почти не снижает вероятность коллизий, но добавляет ещё больше нулей, что становится неэффективно, поскольку делает данные сильно разреженными. На практике длины в 1024 бита обычно достаточно.

Говоря о структурных особенностях химических соединений, стоит упомянуть ещё несколько важных для расчёта дескрипторов. Одним из таких является Ван-дер-Ваальсовый объем молекулы. Кроме него было придумано достаточное количество описаний **поверхности** молекул (поверхность Конноли, поверхность доступная растворителю, полярная поверхность), однако многие методы описания поверхности молекулы используют Ван-дер-Ваальсовы радиусы как основу для построения геометрических моделей. Суть его поверхности заключается в объединении в единую молекуляр-

ную систему радиусов, центрированных на атомах молекулы. Напомним, что В-д-В радиус – это эмпирически определенный для каждого атома радиус на основе анализа кристаллических структур и потенциала Леннард-Джонса [96]. Он представляет собой усредненное расстояние, при котором силы притяжения и отталкивания между атомами уравновешиваются. То есть это «граница» атома, за которую не могут заходить другие атомы, не образуя при этом с ним химическую связь. Чем он больше, тем «пухлее» атом и тем сильнее он может взаимодействовать с соседями. Именно такими сферами и окружена поверхность всей молекулы при расчёте этого дескриптора. Когда она может быть полезна? На самом деле, многие свойства молекул, которые нас интересуют, так или иначе коррелируют с объемом, занимаемым молекулой. Особенно отчетливо это проявляется на стадии докинга (процесс стыковки малой молекулы с одним из сайтов биологической мишени). При моделировании взаимодействия малой молекулы с сайтом биологической мишени важно понимать стерические особенности обоих. Одним из способов визуально оценить, насколько форма лиганда хорошо подходит полости сайта связывания можно с помощью построения поверхности Ван-дер-Ваальса. Кроме того, значения оценочных функций докинга, которые рассчитывают комплементарность взаимодействия молекулы с биологической мишенью, напрямую коррелируют с площадью контакта между поверхностями лиганда и сайта связывания. Чем больше контакт гидрофобных частей – тем лучше аффинность связывания [97].

Стоит сказать, что количество дескрипторов в химической информатике исчисляется тысячами. Их классификация максимально разнообразна. Часто один и тот же дескриптор можно отнести к разным классам. Одним из таких популярных примеров является дескриптор E-State – дескриптор, отображающий топологическое влияние соседей атома на его электронные свойства [98]. Он считается по формуле:

$$Ei = Ii + \Delta Ii,$$

где Ii – собственные электронные свойства атома (электроотрицательность, гибридизация и т.д.), а ΔIi – поправка на влияние соседних атомов.

Такой дескриптор по праву может относиться и к топологическим, поскольку учитывает влияние окружения, и к электронным, поскольку отражает свойства атома. Его применение часто упоминается в контексте прогноза биодоступности, так как высокие значения E-State на кислородах у молекулы коррелируют с улучшенной растворимостью в полярной среде.

Подводя небольшие итоги, можно отметить, что дескрипторы в QSAR-моделировании играют ключевую роль, позволяя переводить сложную молекулярную структуру в набор числовых характеристик, пригодных для статистического и машинного анализа. Их разнообразие (от простых физических параметров до сложных топологических и электронных индексов) обеспечивает гибкость при построении моделей для самых разных задач: от прогноза растворимости и токсичности до оценки сродства к биомишеням.

4.1.5. Построение и валидация моделей

При рассмотрении процесса построения и валидации моделей QSAR важно сразу отметить, что такие модели обучаются на определенных датасетах, которые включают соединения из ограниченного числа регионов химического пространства. Например, в вашем наборе данных могут быть фенольные соединения, алкалоиды и терпеноиды. Тогда следует понимать, что предсказания модели для соединений из этих классов будут более достоверными, чем для других (например, алифатических углеводородов). Это связано с понятием **домен применимости (applicability domain)** – областью химического пространства, в пределах которой модель способна делать надежные прогнозы [99]. Модели, построенные на разнородных или слишком широких наборах, как правило, менее устойчивы и хуже предсказывают свойства новых соединений, сильно выходящих за пределы обучающей выборки.

Домен применимости определяет, для каких молекул модель будет давать корректные прогнозы. Если структура нового соединения существенно отличается от всех молекул обучающей выборки (например, по ключевым дескрипторам или фингерпринтам), то предсказание модели становится ненадежным. Для контроля этого параметра часто используют методы анализа расстояний в пространстве дескрипторов, визуализацию главных компонент (PCA – Principal Component Analysis) [100] или расчёт коэффициента сходства по Танимото (Tanimoto similarity) [101].

Определение домена применимости обычно происходит на завершающем этапе. Само же построение модели на практике включает несколько шагов: сбор и очистка данных, расчёт дескрипторов, их отбор, выбор подходящего алгоритма машинного обучения, само обучение и валидация модели. Каждый из этих этапов требует внимательного подхода, так как ошибки на ранних стадиях могут привести к некорректным или не интерпретируемым результатам.

Важно подчеркнуть, что на стадии расчёта дескрипторов часто происходит первоначальный подсчет большого количества различных характеристик молекул. Во многих инструментах представлен расширенный набор дескрипторов, не все из которых будут достаточно информативными для конкретного исследования модели. Например, в пакете Mordred включено для расчёта более 1800 параметров (от топологических индексов до фармакофорных признаков) [102]. Очевидно, не все из них войдут в конечный набор, поскольку избыточное количество признаков приводит к переобучению модели (overfitting). С другой стороны, раннее утверждение о том, какое конкретное число дескрипторов должно присутствовать в обучающей модели, тоже не является самым верным подходом. Ведь таким образом исследователь лишает себя возможности увидеть неожиданные корреляции между конкретным дескриптором и свойством, о котором раньше мог не догадываться в явном виде. Например, дескриптор, не имеющий очевидной связи со свойством, может оказаться значимым в комбинации с другими признаками. Например, в работе Shen et al. (2004) при

моделировании активности антиконвульсантов дескрипторы, описывающие гибкость молекулярной цепи, неожиданно показали высокую значимость, хотя изначально исследователи сфокусировались на электростатических свойствах [103]. Это подчеркивает важность сохранения широкого набора дескрипторов на этапе предварительного анализа.

В то же время, стоит учитывать те дескрипторы, между которыми наблюдается высокая взаимная корреляционная зависимость или, так называемая, мультиколлинеарность. Мультиколлинеарность – это статистическое явление, при котором два дескриптора или более сильно коррелируют между собой ($|r| \geq 0{,}8$), по сути, дублируя информацию. Это искажает вклад признаков в модель и снижает ее устойчивость. Взаимные корреляции между дескрипторами определяются через матрицу корреляции, где каждый элемент в такой матрице показывает силу связи между парой признаков. В качестве показателя корреляции рассчитывают коэффициент Пирсона для линейных зависимостей или Спирмена для нелинейных [104]. Кроме такой матрицы можно провести VIF-анализ (Variance Inflation Factor) [105]. Он оценивает, насколько дисперсия дескриптора объясняется другими признаками. Принято, если VIF>5-10, то это указывает на сильную мультиколлинеарность. Его главное ограничение состоит в том, что он может улавливать только линейные зависимости между дескрипторами, поэтому использовать данный метод следует в сочетании с матрицами корреляций. При обнаружении дескрипторов со взаимной корреляцией следует либо удалить один из них, либо заменить коррелирующие дескрипторы на их главные компоненты (PCA), либо применить регуляризацию L1/L2 (частично снижает мультиколлинеарность). Кроме того, при расчёте большого количества дескрипторов для разных молекул, некоторые из них имеют нулевые значения, поскольку могли отражать тот признак, которого нет ни в одной молекуле. Такие нулевые дескрипторы однозначно стоит удалять, поскольку никакого вклада (кроме шума) в обучение модели они вносить не будут.

Тогда может возникнуть закономерный вопрос: «Как из общего числа рассчитанных дескрипторов выбрать наиболее информативные?». Экспертные знания (Domain knowledge) в области химии для этой операции являются неотъемлемой частью анализа, исследователь должен быть способен на принятие решения об исключении или добавлении того или иного дескриптора в модель, опираясь на свой опыт, мозговые штурмы и данные из литературы. Существуют и алгоритмические подходы для выбора наиболее информативных признаков. Среди них можно выделить использование feature importances в Random Forest или других методах машинного обучения; LASSO – регрессия, которая обнуляет коэффициенты менее значимых дескрипторов или RFE (Recursive Feature Elimination) – итеративное удаление наименее значимых признаков без значительной потери качества в моделях. С каждым из этих механизмов стоит отдельно ознакомиться в специализированной литературе по машинному обучению [106]. Здесь мы лишь обсуждаем подходы, которые помогут прийти к желаемому результату.

Стоит уточнить, что перед имплементированием дескрипторов в модель обязательно нужно привести значения этих дескрипторов к единому масштабу. Такая стандартизация необходима для того, чтобы ни один из признаков не доминировал над другими только из-за разницы в диапазонах значений. Например, молекулярная масса имеет заметно отличающиеся от остальных дескрипторов большие значения. То есть, если один дескриптор измеряется в тысячах, а другой в сотнях, первый будет автоматически вносить больший вклад в модель, даже если его информативность ниже. Это особенно важно для методов машинного обучения, использующих расстояния между объектами (например, k-NN, SVM), а также градиентные методы и регуляризацию [107].

Из способов стандартизации выделяют:

1) Z - нормализацию:

$$z = \frac{x_i - \mu}{\sigma}$$

где μ – среднее значение одного дескриптора, а σ - стандартное отклонение. В результате после такой нормализации каждый дескриптор будет иметь среднее значение 0 и стандартное отклонение 1.

2) MinMax Scaling: приводит значения к диапазону [0;1].

$$z_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}$$

Стандартизация дескрипторов, впоследствии, позволяет легче сравнивать коэффициенты моделей между собой. К примеру, в той же самой линейной регрессии [108].

После преодоления всех трудностей, связанных с имплементированием дескрипторов в модель, наступает процесс выбора метода модели машинного обучения, с помощью которого будет предсказано то или иное свойство. Фактически, в исследованиях строят сразу несколько типов моделей, качество которых затем оценивают с помощью различных показателей. Для регрессионной задачи используют коэффициент детерминации (R^2), среднеквадратичную ошибку (RMSE) или среднюю абсолютную ошибку (MAE). В случае классификационной задачи значимыми являются показатели точности (Accuracy), чувствительности (Sensitivity), F1-меры или кривой ROC-AUC.

В роли моделей машинного обучения в химической информатике в зависимости от задачи часто выступают: SVM (Support Vector Machine), MLP (Multi-Layer Perceptron), XGB (eXtreme Gradient Boosting), LR (Linear Regression), RF (Random Forest), k-NN (k-Nearest Neighbors), NB (Naive Bayes) и другие [106, P. 4]. Каждый из перечисленных алгоритмов обладает своими преимуществами и ограничениями, что определяет их применимость к различным типам данных и задач. Например, линейная регрессия и логистическая регрессия

хорошо подходят для задач, где между дескрипторами и целевым свойством предполагается линейная зависимость. В то же время, ансамблевые методы, такие как случайный лес и градиентный бустинг, способны выявлять сложные нелинейные закономерности и зачастую демонстрируют высокую устойчивость к переобучению даже при большом количестве признаков.

Многослойный перцептрон (MLP), относящийся к классу искусственных нейронных сетей, отличается способностью аппроксимировать практически любые зависимости, однако требует тщательной настройки архитектуры и значительных вычислительных ресурсов. Метод опорных векторов (SVM) эффективен при работе с высоко размерными пространствами признаков и может использовать различные ядровые функции для учета нелинейных связей. Алгоритм k-ближайших соседей (k-NN) прост в реализации, однако чувствителен к масштабу данных и размеру обучающей выборки. Наивный байесовский классификатор (NB) отличается высокой скоростью работы, но предполагает независимость признаков, что не всегда соответствует реальным данным в химии [109].

В процессе обучения любой из этих моделей критически важным шагом становится ее валидация, которая позволяет оценить предсказательную способность и устойчивость алгоритма к переобучению. Важно понимать, что высокая точность на обучающей выборке не гарантирует успешное применение модели к новым данным. Для того, чтобы избежать некорректного использования информации из исходной выборки соединений, используют различные методы разбиения анализируемого датасета.

Одним из таких подходов является разделение данных на обучающую (60-80%), валидационную (10–20%) и тестовую (10–20%) выборки [110]. Обучение проводится на первой части, гиперпараметры настраиваются на второй, а финальная оценка происходит на третьей части, что исключает «подгонку» под тестовые данные.

Другим наиболее популярным методом является кросс-валидация (5-кратная, чаще 10-кратная). В этом методе данные разбиваются на k частей, модель обучается на k−1 частях, а проверяется на оставшейся. Процесс повторяется k раз для каждого набора, а результаты усредняются. Этот подход особенно полезен при малом объеме данных, что является довольно частой проблемой у разного рода исследованиях в химической информатике [111].

Таким образом, построение и валидация QSAR-моделей – это многоэтапный процесс, требующий сочетания современных вычислительных методов, глубокого понимания химических особенностей объектов и строгой оценки результатов. Только комплексный подход, включающий тщательную работу с данными, выбор оптимальных алгоритмов и анализ интерпретируемости, позволяет создавать надежные и практически значимые модели для прогнозирования свойств химических соединений. Если вам интересно погрузиться в более технические детали построения моделей и их валидации, рекомендуем к прочтению 3 том Т.И. Маджидова и коллег «Введение в хемоинформатику. Моделирование «структура-свойство»» [112].

4.2. Виртуальный скрининг

Как найти иголку в стоге сена?

После успешного построения и валидации QSAR-моделей следующим важным этапом в процессе разработки новых химических соединений и лекарственных средств становится виртуальный скрининг (Virtual Screening, VS) [113]. Это процесс, при котором отбираются перспективные молекулы из огромных баз данных, сокращая время и затраты на экспериментальные исследования. Он является альтернативой HTS [114].

Виртуальный скрининг – это вычислительный подход, направленный на быстрое и экономичное выявление потенциально активных соединений, обладающих заданными свойствами, например, биологической активностью, низкой токсичностью или подходящей фармакокинетикой из большого количества разных соединений. Он служит мостом между химической информатикой и экспериментальной химией, позволяя исследователям сосредоточиться на синтезе наиболее перспективных кандидатов.

В предыдущей главе мы рассмотрели методы построения и валидации моделей QSAR, которые являются одним из ключевых инструментов при виртуальном скрининге. Используя предсказательные модели, можно оценивать свойства тысяч и даже миллионов соединений, отбирая те, которые с высокой вероятностью обладают желаемыми свойствами. Однако такие модели являются не единственным способом, при котором оценивается потенциальная способность молекулы стать важной находкой для поставленной задачи.

Виртуальный скрининг, в том числе, может быть основан на фармакофорном поиске – подборе молекул под специальную 3D модель ключевых признаков в соединении, удовлетворяющих поисковым условиям. То есть под фармакофором мы будем понимать объединение в единую пространственную модель ключевых стерических и электронных признаков химических соединений (рис. 39), которые ответственны за проявление того или иного свойства в молекулах (чаще – биологическую активность) [115]. Такие модели могут быть построены концептуально двумя разными способами.

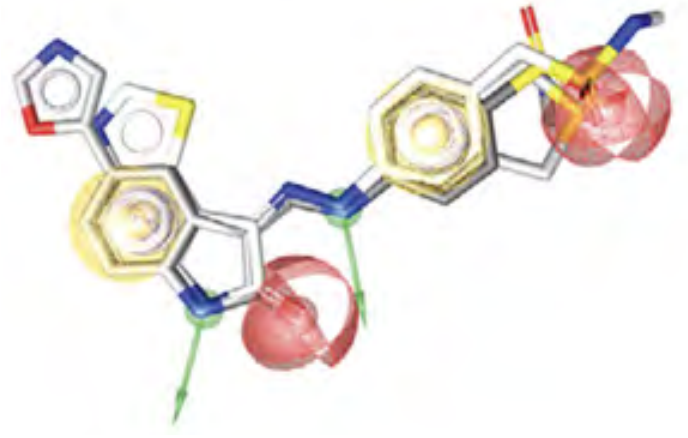


Рис. 39. Фармакофорная модель лиганда. Жёлтыми сферами показаны гидрофобные признаки в молекуле, красными сферами показаны признаки акцепторов водородной связи, синие сферы отображают признаки доноров водородной связи [116]

Наиболее популярным и давно устоявшимся рабочим методом является построение Ligand-based Pharmacophore, то есть фармакофора, который получен на основе информации об известных лигандах с необходимыми свойствами. В таком методе известные структуры активных молекул выравниваются друг относительно друга и выделяются их общие структурные закономерности в пространстве: доноры и акцепторы водородных связей, гидрофобные фрагменты, фрагменты π-π взаимодействий и т.д. Из этих признаков, расположенных в пространстве с конкретными координатами получается 3D модель, которую в дальнейшем используют в виртуальном скрининге при анализе библиотеки на предмет подходящих химических соединений [117].

Другим, менее популярным, но многообещающим методом является построение фармакофорных моделей лишь на основе знаний структуры сайта связывания биологической мишени (актуально, если речь идет о поиске биологически активных соединений). Такие методы появились в связи со стремительным ростом количества оцифрованных 3D структур различных белков, которые не удавалось получить раннее. Преимущественно эти методы используют докинг различных малых органических фрагментов в сайт связывания мишени и дальнейший выбор наиболее значимых признаков (мест, где связались эти фрагменты) на основе оценочной функции докинга и других показателей. Такие фармакофоры тоже имеют трехмерную модель и могут быть аналогичным образом использованы для поиска необходимых соединений [117].

Кроме фармакофорного поиска имеет место в виртуальном скрининге и поиск по молекулярному подобию [118]. Это процесс, при котором с помощью различных метрик сходства выделяют близкие структурно молекулы, ожидая, что они будут обладать схожими свойствами. Одной из таких популярных метрик является коэффициент сходства по Танимото.

$$T = \frac{N_c}{N_a + N_b - N_c},$$

где N_a– количество элементов во множестве A, N_b – количество элементов во множестве B, N_c – количество элементов во множестве C.

Коэффициент Танимото измеряет степень схожести двух множеств. Это достаточно удобная формула, которая принимает значение нуль, если во множествах нет ни одного пересекающегося элемента, и единицу, если все элементы множеств совпадают. Для анализа молекулы разбиваются на структурные фрагменты (например, с помощью уже известных нам фингерпринтов) и их множества сравниваются друг с другом [101].

Также эффективна при виртуальном скрининге обычная работа с известными в базе данными: фильтрация по подходящим фармакокинетическим или фармакодинамическим свойствам, молекулярной массе, применение различных ТОХ-фильтров и т.д [119].

Стоит сказать, что при полноценном виртуальном скрининге эти методы комбинируются для проведения более качественного анализа базы данных и выявления необходимых молекул. Это многоэтапный процесс, объединяющий методы хемоинформатики, машинного обучения и молекулярного моделирования. Его эффективность зависит от качества исходных данных, выбора стратегии и интеграции с экспериментальными исследованиями [113].

4.3. Ретросинтез

Синтез мочевины, осуществленный Фридрихом Велером в 1828 году, стал отправной точкой в развитии органического синтеза [120]. С тех пор подходы к синтезу, сложность синтезируемых молекул и реактивы для этих целей претерпели немало изменений. Химия с годами становится всё сложнее, учёные исследуют химическое пространство всё больших по размеру молекул, открывают новые реакции и стратегии синтеза. Эту эволюцию несложно проследить даже на примере лекарственных препаратов, которые были открыты ранее и тех, которые появляются на рынке сейчас. На рисунке 40 приведены структуры аспирина (2-ацетилоксибензойная кислота) и дабрафениба (N-[3-[5-(2-аминопиримидин-4-ил)-2-трет-бутил-1,3-тиазол-4-ил]-2-фторфенил]-2,6-дифторбензолсульфонамида).

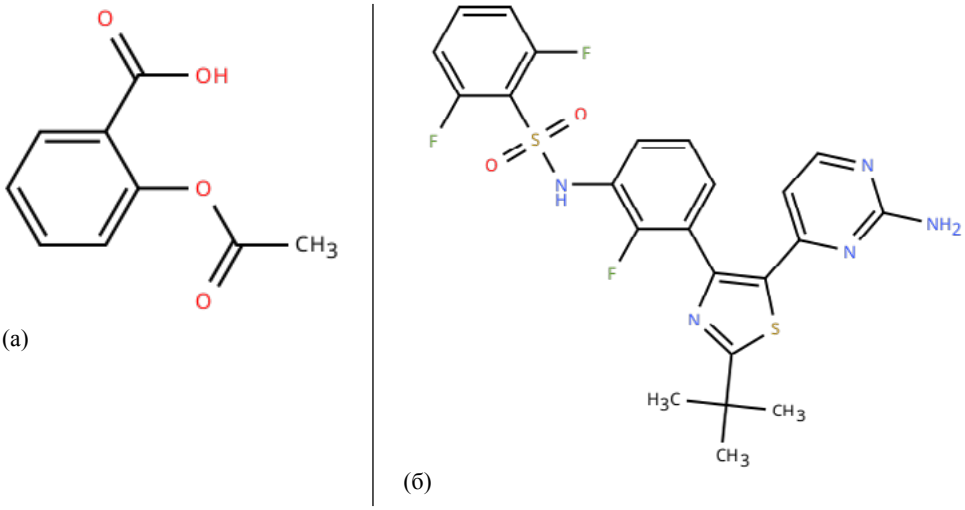


Рис. 40. Структуры аспирина (а) и дабрафениба (б)

В первом случае аспирин – давно известный человечеству нестероидный противовоспалительный лекарственный препарат. Во втором случае, дабрафениб – это современный таргетный противоопухолевый препарат из класса ингибиторов киназ, применяемый для лечения меланомы, немелкоклеточного рака легкого и анапластического рака щитовидной железы, вызванных специфической мутацией BRAF V600 [121]. Его сложная молекула (контрастирую-

щая с простой структурой аспирина) была разработана для точечного воздействия на конкретную молекулярную мишень (мутантный белок BRAF) внутри раковой клетки, что иллюстрирует переход от эмпирической терапии к высокоспецифичной, основанной на понимании молекулярных механизмов болезни и персонализированной медицине [122]. Этот пример демонстрирует, что сложность «полезных» для общества органических молекул с годами только растет и не похоже, что тенденции будут меняться.

В этом контексте важно отметить, что чем разнообразнее и больше структура молекулы, тем, соответственно, ее сложнее синтезировать. Даже опытный химик, смотря на сложносочиненную молекулярную структуру, не всегда сможет дать ответ на вопрос о том, как ее синтезировать оптимальным образом. Под оптимальным путем мы будем понимать максимально возможный выход, минимальную стоимость затрат на реактивы, небольшое количество стадий в синтезе и, конечно же, полученный результат в виде целевой молекулы без лишних примесей. Такое большое количество параметров, которое нужно учитывать при планировании синтеза просто технически сложно удержать в голове, поскольку комбинаторика процесса слишком велика. Именно в этот момент и появляется помощник в виде алгоритмов машинного обучения.

К настоящему времени разработаны несколько подходов, благодаря которым получается адаптировать эти алгоритмы для применения в планировании синтеза сложных молекул, расчёте его стоимости и оценки качества предложенных путей [123]. Стоит отметить, что для сложных органических молекул в начале исследований часто используют ретросинтетический подход.

Ретросинтез – это метод в органической химии, используемый для проектирования синтеза сложных молекул. Он заключается в анализе целевой молекулы и её мысленном «разделении» на более простые компоненты путем разрыва ключевых химических связей [124]. Последовательное упрощение структуры Target Molecule (TM) в соответствии с определенными правилами проводят до тех пор, пока не будет получено доступное соединение, либо такое соединение, способ синтеза которого известен. Этот обратный подход позволяет определить, из каких доступных исходных веществ и с помощью каких реакций можно собрать целевую структуру.

В настоящее время всего несколько передовых ИИ платформ в области органической химии имеют встроенные в свои модели алгоритмы, позволяющие прогнозировать ретросинтез разных органических молекул. Одной из таких платформ является Синтелли [41]. Этот инструмент позволяет спрогнозировать до 5 схем многостадийного синтеза малой органической молекулы с помощью нейронных сетей. Платформа будет строить ретросинтетическое дерево до достижения коммерчески доступных молекул или до достижения лимита в 5 стадий. Прогнозируемые ретросинтетические пути также имеют оценку уверенности модели. На рисунке 41 приведён пример ретросинтетического анализа молекулы Гефитиниба – онкологического лекарственного препарата, применяемого в терапии рака легкого. Препарат был разработан компанией AstraZeneca и распространяется под торговым названием Иресса [125].

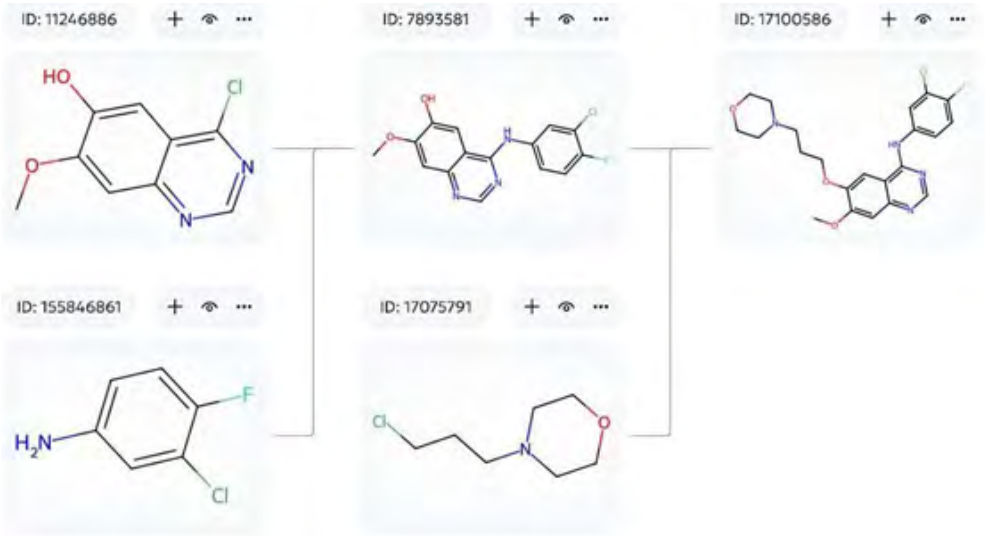


Рис. 41. Пример ретросинтеза молекулы Гефитиниба на платформе Синтелли [41]

Этот пример хорошо показывает работу платформы в действии. Все конечные реагенты, из которых синтезируется Гефитиниб (самая правая молекула на рисунке) коммерчески доступны. Кроме того, говоря о коммерческой доступности, на данной платформе можно спрогнозировать стоимость самого синтеза. В качестве примера приведём расчёт стоимости последней стадии синтеза Гефитиниба при необходимости получить 100 г конечного продукта (рис 42).

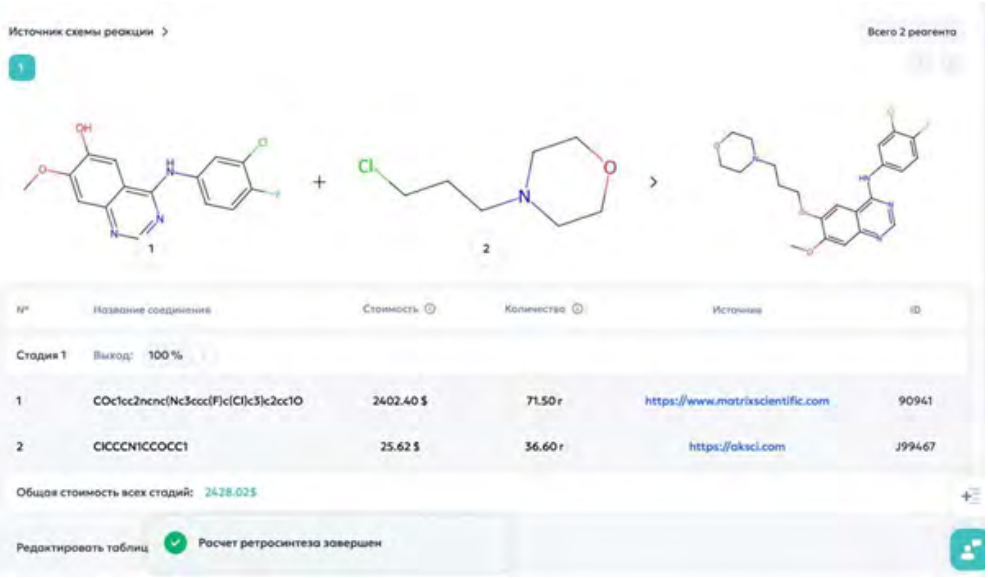


Рис. 42. Расчёт стоимости финальной стадии при синтезе Гефитиниба на платформе Синтелли [41]

Однако стоит понимать, что платформы вроде Синтелли выступают не как замена эксперту, а как мощный «когнитивный усилитель», расширяющий возможности химика-синтетика. Они берут на себя рутинный перебор вариантов и первичную оценку, позволяя исследователю сосредоточиться на творческих аспектах дизайна синтеза, оптимизации выбранного пути и решении действительно сложных задач, где требуется глубокое понимание механизмов реакций и стереохимии.

Если вас заинтересовала информатика реакций, их представление и прогнозирование путей ретросинтеза и синтеза, рекомендуем к прочтению 5 том Т.И. Маджидова и коллег «Введение в хемоинформатику. Информатика химических реакций» [126].

5. ПРИМЕРЫ ПРАКТИЧЕСКИХ ПРИЛОЖЕНИЙ

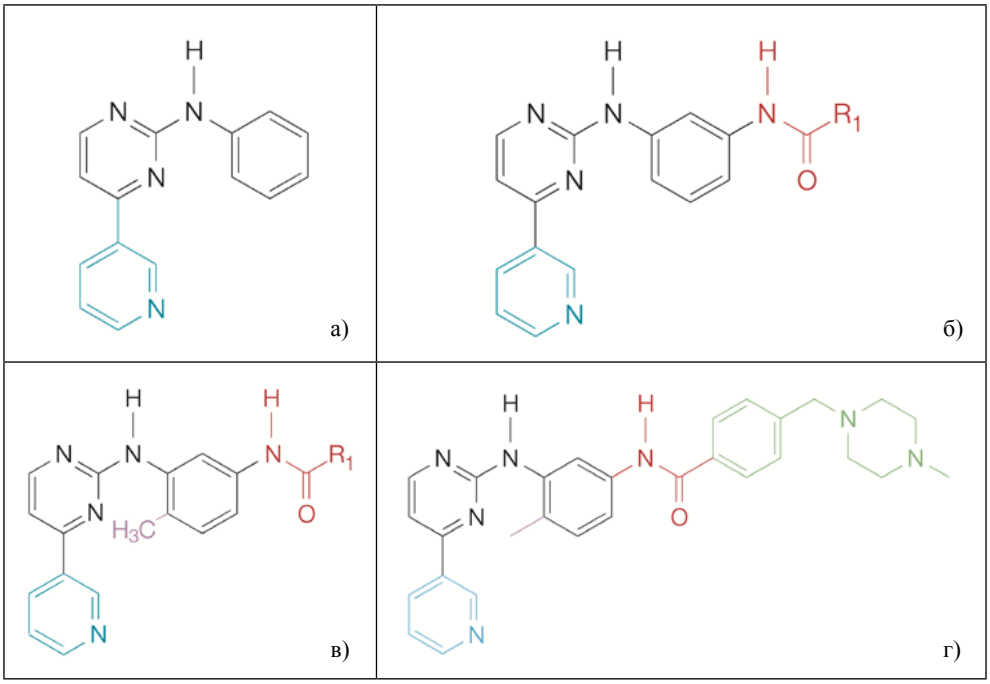
5.1 Использование хемоинформатики в фармацевтической индустрии

На протяжении всей книги мы говорили о том, что химическая информатика стала незаменимым инструментом для учёных в разработке лекарств, особенно в контексте рационального дизайна молекул. Под рациональным дизайном мы будем иметь в виду обоснованное конструирование химических соединений с заданной биологической активностью и необходимыми ADMET свойствами, а также дальнейшую оптимизацию их структуры [127].

Представьте, что Вы архитектор, проектирующий дом. У Вас есть четкий план: количество комнат, материалы, расположение окон. Но вместо чертежей – молекулы, вместо кирпичей – атомы, а вместо пожеланий заказчика – требования к безопасности и эффективности. Именно так работает рациональный дизайн в фармацевтике, где химическая информатика выступает в роли умной линейки или микроскопа. Всего полвека назад открытие лекарств напоминало лотерею. Учёные синтезировали тысячи соединений во многом полагаясь на общие принципы, экспертное знание и метод «проб и ошибок» (относительно случайных), тестировали полученные соединения на активность и надеялись на удачу. Сегодня химическая информатика позволяет во многих случаях заменить случайность логикой. Например, алгоритмы молекулярного докинга показывают, как потенциальный препарат «вписывается» в мишень, словно ключ в замок [128]. А QSAR-модели предсказывают, не станет ли молекула токсичной для печени, ещё до того, как ее синтезировали [129].

В качестве реального примера того, как можно дойти от идеи до препарата с помощью методов химической информатики и рационального дизайна структуры можно привести препарат Иматиниб. Это ингибитор тирозинкиназы, который был разработан для лечения хронического миелоидного лейкоза. Соединение-лидер – производное 2-фениламинопиримидина было установлено с помощью высокопроизводительного скрининга потенциальных ингибиторов протеинкиназы-С (PKC) [91]. Далее, с помощью моделей SAR и докинга был принят ряд ключевых решений об оптимизации структуры для достижения желаемых свойств (табл. 8)

Таблица 8. Процесс оптимизации структуры соединения лидера ингибитора тирозинкиназы. а) Добавление 3′-пиридильной группы (синий) в 3′-положение пиримидина усиливало клеточную активность; б) Амидная группа (красная), присоединенная к фенильному кольцу, обеспечивала активность против тирозинкиназ; с) Метил (фиолетовый), присоединенный к диаминофенильному кольцу, увеличивал селективность действия против тирозинкиназ; д) Окончательное присоединение N-метилпиперазинового кольца (зеленый) заметно повысило растворимость соединения и его пероральную биодоступность [130].



Иматиниб является ярким примером того, как комбинация HTS, SAR, молекулярного докинга и структурного анализа позволила создать «умную» молекулу, изменившую подход к лечению онкологического заболевания. Однако, это не единственный пример успешного применения химической информатики в рациональном подходе при разработке лекарств.

Препарат, направленный на лечение легочной гипертензии Бозентан, был разработан с использованием методов фармакофорного моделирования и QSAR моделей для оптимизации липофильности и биодоступности [131].

Саксаглиптин – антидиабетический препарат, ингибитор DPP-4 (дипептидилпептидазы-4) для лечения диабета 2-го типа, был найден при виртуальном скрининге и последующем докинге в сайт DPP-4. Молекулярная динамика подтверждала стабильность комплекса лиганд-белок. Результат – саксаглиптин одобрен FDA в 2009 году и широко применяется в комбинированной терапии диабета [132].

В 2023 году компания Insilico Medicine сообщила об успешном прохождении доклинических исследований молекулы INS018-055 для лечения идиопатического легочного фиброза. Эта молекула была сгенерирована с помощью искусственного интеллекта, который одновременно идентифицировал и терапевтическую мишень, и потенциальный кандидат на лекарство. Молекула успешно завершила I фазу клинических исследований, подтвердив безопасность применения [133].

Все эти примеры служат вдохновением для вычислительных химиков разрабатывать новые современные методы и более точные модели для предсказания химических свойств различных молекул. Хемоинформатика стала неотъемле-

мой частью современного фармацевтического R&D (research and development - исследований и разработок), позволяя значительно сократить время и стоимость разработки новых лекарственных средств, повысить качество кандидатов и уменьшить количество неудач на поздних этапах испытаний, связанных с недостаточно оптимизированными свойствами ADMET. С развитием искусственного интеллекта и вычислительных технологий роль хемоинформатики в фармацевтической индустрии будет только расти [2].

5.2. Экологическая химия

Экологические вопросы в современном мире интересуют всё большее количество людей с каждым годом. Это не удивительно, ведь окружающая нас среда напрямую влияет на состояние нашего здоровья и отражается на следующих поколениях. Тысячи химических веществ, созданных человеком, накапливаются в почве, воде и воздухе, угрожая экосистемам. Не говоря уже об экологических катастрофах (разлив нефти, пожары на мусорных свалках, утечка отходов на заводах и т.д.), происходящих каждый год с некоторой периодичностью. Всё это происходит очень быстро, в современных индустриальных условиях, где производство набрало колоссальные масштабы и скорости, сложно оградить себя от постоянного воздействия тех или иных химических веществ. Мы пьем воду из бутылок, в которых содержится микропластик. Мы употребляем в пищу продукты, которые проходят целый комплекс химических обработок перед продажей ради усиления вкуса и увеличения их срока годности. Мы ездим на машинах, которые выбрасывают миллиарды тонн выхлопных газов каждый год. Мы выстраиваем производство товаров, отходы от которого, впоследствии, попадают в сточные воды. Безусловно, всё это не может пройти мимо нашего организма, пагубно влияя на него в большинстве случаев.

Применение химической информатике нашлось и в данном контексте. Поскольку предсказывать можно любые свойства, подбирая подходящие параметры и условия, экологические аспекты в тех же QSAR моделях могут быть успешно учтены. К примеру, соединения из группы бензотриазолов десятилетиями использовались как антикоррозийные добавки в авиационном топливе. Однако их высокая устойчивость к разложению привела к загрязнению водоёмов. В проекте CADASTER (ЕС) хемоинформатики разработали QSAR-модели, предсказавшие экотоксичность более 300 бензотриазолов для водорослей и рыб. Это позволило идентифицировать наиболее вредные соединения и заменить их на безопасные аналоги [134]. Кроме того, с помощью компьютерного моделирования можно прогнозировать распространение вещества, время полураспада и эффективные способы его нейтрализации в различных средах. В 2023 году японские учёные использовали алгоритмы для подбора растворителей в реакции Сузуки. Модель учитывала не только эффективность реакции, но и стоимость утилизации отходов, а также углеродный след производства. В результате правильного подбора оптимальный растворитель сократил выбросы CO_2 на 70%, а затраты – на 67% [135].

В качестве ещё одного примера можно привести процесс биодеградации сложных загрязнителей. Перфторалкильные вещества (PFAS) - «вечные химикаты», которые десятилетиями накапливаются в организме. Их удаление - глобальная проблема. Учёные из США создали биомиметическую платформу RAPIMER, сочетающую адсорбцию PFAS на растительных наноматериалах и последующее разложение грибком *Irpex lacteus*. Хемоинформатика помогла спрогнозировать, какие фрагменты PFAS наиболее уязвимы для грибковых ферментов и оптимизировать структуру адсорбента для захвата максимального объема загрязнителей. Как итог – система очищает воду от PFAS на 98% за 24 часа [136].

Как мы уже успели убедиться на других примерах, создание каждой новой базы данных всегда сопровождается позиционированием, которое бы отличало ее от предыдущих альтернатив. Поскольку химическое разнообразие огромно (структуры, свойства, эксперименты), специфические показатели, которые не всегда отражены в привычных базах данных, требуют систематизации для цифровых или других исследований. Одним из таких наиболее важных показателей для экологической химии является токсичность. Несмотря на то, что до сих пор не создано полноценного международного ресурса, который бы на основе единого *общепризнанного* стандарта предоставлял информацию обо *всех* экологически важных параметрах молекул, составляющих глобальное химическое пространство, существует несколько специализированных баз данных, которые аккумулируют в себе специализированную информацию о разнообразии токсических эффектов (доза, конечное воздействие на организм, длительность эффекта и т.д.), которыми может обладать та или иная молекула. Одной из таких наиболее популярных баз данных является TOXRIC. Это открытая база данных, в которой содержится информация о более чем 113 тысячах химических соединений, охарактеризованных по 13 категориям токсичности и включающих около 275 параметров (endpoints) и 38 типов дескрипторов, таких как структурные, транскриптомные и метаболические данные [137]. Такая комплексность позволяет использовать базу для построения моделей количественной зависимости структура–активность (QSAR) с применением методов машинного обучения, что значительно ускоряет и упрощает процесс оценки токсичности веществ без необходимости проведения дорогостоящих и длительных экспериментов *in vivo* или *in vitro*. Основное назначение TOXRIC – служить источником данных для построения QSAR моделей, что важно для оценки рисков химических веществ в окружающей среде и контроля безопасности промышленных химикатов. Благодаря структурированным и стандартизированным данным TOXRIC способствует развитию вычислительной токсикологии и химической информатики, позволяя исследователям создавать более точные и надёжные модели предсказания токсичности. Может показаться, что информация о 113 тысячах соединений по сравнению с миллионами в обсуждаемых выше базах данных это совсем небольшое количество. На самом деле, в контексте измерения показателей токсичности это отличный по современным меркам результат. Токсикологические исследования дорогие, требуют

много времени и ресурсов (накормить мышку, оценить эволюцию токсических эффектов во времени и т.д.), именно поэтому даже самая маленькая база данных имеет большую ценность. О важности таких баз данных мы ещё поговорим в следующих главах.

5.3. Пищевая промышленность

Химическая информатика, как мы уже поняли, может быть задействована во многих областях, смежных с фармацевтической. Несмотря на то, что первые шаги в этой науке были связаны именно с индустрией разработки лекарств, на данный момент ширина покрытия ее применимости сильно возросла. Пищевая промышленность является той частью химии, в которой новые разработки и предсказания не просто важны, а жизненно необходимы для обеспечения здоровыми продуктами потребителей разных национальностей. Конечно же предметом анализа, различных предсказаний и разработок являются продукты питания. Пища – это сложная смесь компонентов: воды, жиров, белков, витаминов, минералов и множества добавок. Последние играют ключевую роль в сохранении качества, вкуса и безопасности продуктов. Например, антиоксиданты замедляют окисление, консерванты предотвращают порчу, а красители придают привлекательный вид. Все они должны быть зарегистрированы, по ним должны создаваться цифровые базы данных с описанием их свойств и регулироваться соответствующими органами. Не даром уже знакомая нам организация FDA имеет в своем названии литеру F, означающую Food. Food and Drug Administration (FDA) одобряет к использованию не только лекарственные препараты, но и различные пищевые добавки, ведь они так же, как и лекарства, оказываются внутри нашего организма, подвергаясь распределению, метаболизму и т.д. Согласно FDA, пищевой добавкой считается любое вещество, которое может стать частью продукта. Однако их безопасность требует строгой проверки. До 1958 года статус «общепризнанных безопасными» (GRAS – Generally Recognized As Safe) присваивался веществам на основе их исторического использования. Сегодня для подтверждения статуса GRAS необходимы научные данные, включая опубликованные исследования и экспериментальные доказательства. Эти данные стали основой для исследований в этой области, позволяя анализировать связи между структурой молекулы и ее безопасностью как пищевого агента [138, P. 2].

Как и в случае с лекарственными препаратами, безопасность пищевой добавки определяется дозой, которая попадает в организм человека. Почему одни молекулы безопасны, а другие нет? Ответ опять же кроется в их химическом строении. Например, лимонная кислота (E330) имеет простую структуру, которая легко метаболизируется организмом, а искусственные красители могут содержать ароматические кольца, связанные с аллергическими реакциями или другими побочными эффектами. Таким образом, пищевая промышленность сталкивается с уникальными вызовами: необходимость обеспечения безопасности продуктов, сокращение использования вредных добавок, повышение

питательной ценности и продление срока годности. В мире занимаются этим вопросом и написано несколько литературных трудов, ключевым из которых является работа Martinez-Mayorga K. и Medina-Franco J. L.: «Foodinformatics: applications of chemical information to food chemistry» [138, P. 2].

Результатом оцифровки этого направления стали базы данных, которые включают в себя разную информацию, касающуюся химических веществ, пригодных или, наоборот, непригодных для использования в качестве пищевых добавок. Так, например, база данных BitterDB содержит в себе информацию о свойствах более чем 2400 веществ, обладающих горьким вкусом и, соответственно, ассоциированных с рецепторами T2Rs и TAS2Rs [139]. Или база данных SuperSweet, которая хранит информацию о более чем 8000 натуральных и искусственных подсластителях, включая их свойства, такие как 3D-структура, происхождение, сладость, одобрение, калорийность и т.д., и предоставляет гипотезы об их связывании с рецептором [140]. Есть и не такие специфические базы данных, основанные на более общей информации о пищевых добавках. Примером таковой является база данных Research Institute for Fragrance Materials (RIFM)/FEMA Fragrance and Flavor, которая содержит информацию о более чем 5000 материалах. Эта коммерческая база данных, она представляет собой один из наиболее полных ресурсов, поскольку содержит информацию о химической структуре (например, номера CAS и SMILES-представления), физико-химических свойствах, синонимах, а в некоторых случаях даже о медицинских и экологических исследованиях [141].

Благодаря накоплению такой информации стало возможным анализировать химическое пространство пищевых добавок. И вот какими результатами мы обладаем на данный момент времени. В исследовании [142] был проведён комплексный анализ обновленного списка FEMA GRAS, включающего 2244 ароматических соединения. Их структура и свойства сопоставлялись с известными человечеству данными о лекарственных препаратах. База из 1713 одобренных FDA лекарств от DrugBank, две коллекции природных химических соединений от разных поставщиков (2449 / 467 молекул) и стандартная библиотека из 10 000 структур, применяемая в высокопроизводительном скрининге. Сравнивали такие показатели химических структур, как липофильность, размер молекул, наличие циклов и др. В качестве дескрипторов были использованы структурные фингерпринты MACCS keys (166-битная версия) и радиальные фингерпринты в коммерческом пакете Canvas от Schrodinger. В качестве методов статистического анализа были выбраны Бокс-Диаграммы для визуализации распределения свойств, анализ главных компонент (PCA) – картирование химического пространства, самоорганизующиеся (SOM) карты – выявление кластеров. В результате анализа получилось сделать несколько важных выводов:

1) Липофильность веществ из FEMA GRAS близка к профилю лекарственных препаратов, что указывает на их потенциальную биодоступность.

2) Размер молекул ароматических веществ в списке GRAS в среднем меньше, чем в других базах данных

3) Некоторые соединения GRAS занимают области химического пространства, характерные для фармацевтических молекул

4) Структурное разнообразие списка сопоставимо с библиотеками лекарств, природных соединений и скрининговых коллекций.

Выводы данного исследования не могут не натолкнуть на мысль о том, что химические пространства пищевой и фармацевтической отрасли сильно пересекаются. Во многом это должно определяться биодоступностью и тех и других [142].

Кроме того, благодаря появлению цифровых баз данных в пищевой отрасли стало возможным построение аналогов моделей QSPR, которые используются в фармацевтической индустрии. В пищевой отрасли их название звучит следующим образом: «Structure – Flavour Relationship». То есть связь химической структуры молекулы с ее вкусовыми и ароматическими характеристиками [142]. Во многих работах, посвященных этой тематике, учёные пытаются выявить подходящие дескрипторы молекул, с помощью которых они могли бы описывать молекулы и объяснять их вкусовые и ароматические свойства. И, конечно, они сталкиваются ровно с теми же проблемами, которые давно известны хемоинформатикам, занимающимся разработкой лекарств. Например, в исследовании [143] при анализе данных для построения моделей, обсуждали явление уже знакомых нам с вами «скал активности», только в пищевой отрасли это называется «скалами вкуса / запаха». Принцип тот же, при достаточно высокой структурной схожести, молекулы могут обладать совершенно разными вкусовыми и ароматическими характеристиками. Это часто обуславливается действием на разные вкусовые рецепторы, что в свою очередь сильно напоминает селективное действие лекарственных препаратов на определенные мишени. Так, авторы вышеупомянутого исследования приводят в пример две молекулы: стерические R- и S- изомеры карвона – природного вещества из семейства терпеноидов (рис 43).

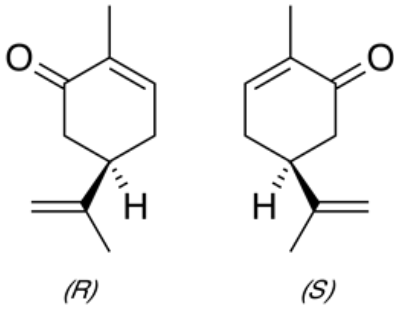


Рис. 43. Структуры R- и S- изомеров карвона

Один из них (R-изомер) обладает ярко выраженным мятным запахом, в то время как другой (S-изомер) имеет запах семян тмина и укропа.

Все эти примеры доказывают, что последовательность действий при поиске химических веществ с заданными свойствами аналогична в разных областях и

сталкивается с похожими вызовами. То же самое происходит в косметической промышленности и в материаловедении. Применение методов химической информатики позволяет на ранних стадиях выявлять необходимые модификации в молекулах без проведения дорогостоящих экспериментов, а также тестировать гипотезы в рамках каждой разработки.

5.4 Хемоинформатика и окружающий мир

5.4.1. Химическое пространство арбуза

Хотелось ли Вам когда-нибудь изучить химическое пространство арбуза? В действительности это интересный вопрос. Догадываетесь ли Вы, как много низкомолекулярных химических соединений человек получает после употребления арбуза? Кроме того, какой природы эти соединения?

К счастью, мы можем теперь исследовать эти вопросы с помощью хемоинформатики. Для этого нам прежде всего нужен датасет соединений, содержащихся в арбузе. Его можно найти в базе данных Coconut [144] в разделе «Collections» под названием «Watermelon». Датасет содержит 710 соединений в формате SMILES.

Теперь нам нужен инструмент, который позволяет визуализировать химическое пространство датасета. Обратимся к платформе Синтелли, к модулю SynMap [41]. Загружаем наш датасет и получаем соответствующую россыпь точек (рис. 44).

Рис. 44. Визуализация датасета ‘Watermelon’
с помощью SynMap модуля платформы Синтелли [41]

Но что всё это значит? Каждая точка соответствует молекуле и имеет определенные координаты. Оси x,y не имеют физического смысла, однако структурно схожие соединения располагаются на карте рядом друг с другом. Кроме того, в некоторых местах мы наблюдаем скопления точек – кластеры. Давайте посмотрим какие соединения входят в их состав. На рисунке 40 представлены по 2 представителя кластеров А, Б, В. Присмотревшись, мы можем сделать вывод, что соединения внутри каждого кластера, действительно, относятся к одним и тем же классам. Так, кластер А – фенольные соединения (производное халкона, оксикоричной кислоты); Б – алифатические аминокислоты (в частности, оксокислоты), В – гликозиды. Таким образом, благодаря инструментам визуализации химического пространства можно довольно быстро получить представление о качественном составе датасета (рис. 45).

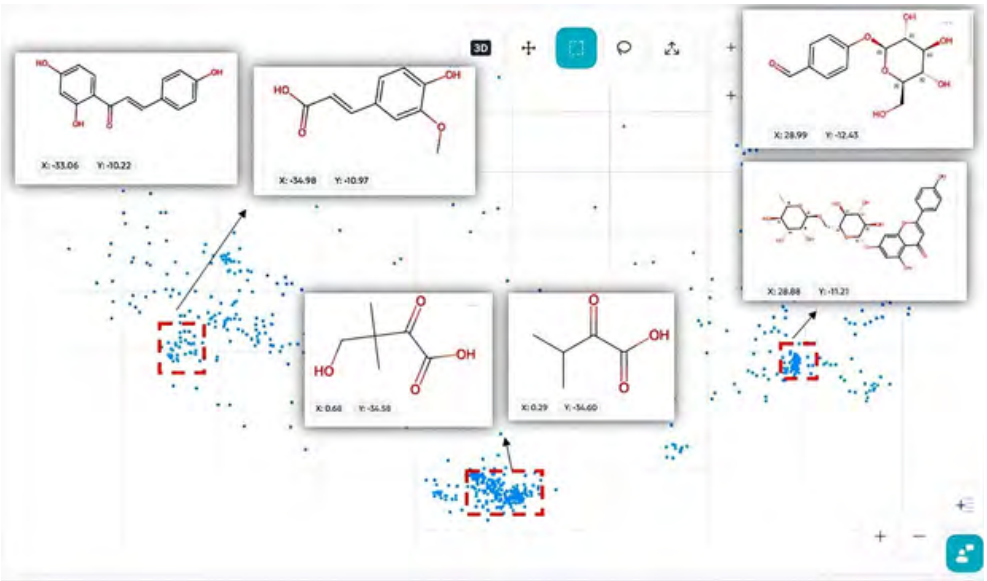


Рис. 45. Кластеры в датасете ‘Watermelon’ в модуле SynMap платформы Синтелли [41]

Строго говоря, описанная процедура относится к визуализации химического пространства. Для изучения теоретических основ и соответствующих инструментов в реализации этого процесса рекомендуем к прочтению 6 том Т.И. Маджидова «Введение в хемоинформатику. Химическое пространство и виртуальный скрининг» [145].

5.4.2. Нитрозамины в лекарственных препаратах: актуальность проблемы и роль хемоинформатики в её решении

В последние годы глобальные регуляторы (FDA) столкнулись с массовыми отзывами препаратов из-за обнаружения нитрозаминов – канцерогенных примесей, образующихся в процессе синтеза или хранения лекарств. Эти сое-

динения, такие как N-нитрозодиметиламин (NDMA) и N-нитрозодиэтиламин (NDEA), классифицируются как вероятные канцерогены. Их присутствие выявлено в популярных препаратах: Сартаны (антигипертензивные средства), Ранитидин (блокатор H_2-рецепторов) и Метформин (сахароснижающий препарат). Росздравнадзор подтвердил, что в 2018 году в китайской субстанции обнаружили NDMA.

Из-за обнаружения канцерогена NDMA в сартанах в России изъяли 561 серию лекарств, а затем – ещё 612 серий 10 препаратов (всего 25 млн упаковок). Параллельно FDA отозвало 233 тыс. флаконов антидепрессанта дулоксетина (Cymbalta) из-за превышения допустимого уровня канцерогена N-нитрозо-дулоксетина. Несмотря на применение с 2004 года, только в декабре 2024-го препарат получил класс риска II – из-за потенциальной опасности нитрозаминов, повышающих риск рака при длительном воздействии.

На рисунке 46 представлен предполагаемый механизм образования нитрозамина в ранитидине [146]. Известно, что монохлорамин (NH_2Cl) при кислом pH подвергается диспропорционированию и гидролизу с образованием дихлорамина ($NHCl_2$) и гипохлорита (HOCl). Эти соединения являются более сильными окислителями и могут ускорять разложение ранитидина, способствуя образованию нитрозаминов.

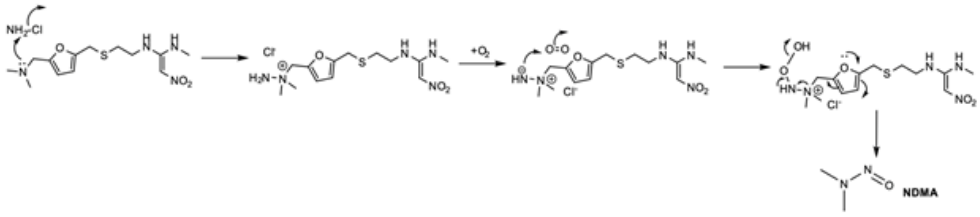


Рис. 46. Предполагаемый механизм образования нитрозамина в ранитидине [146]

С помощью платформы Синтелли возможно также предположить протекание побочных реакций, например ранитидина с хлорамином (Рисунок 47):

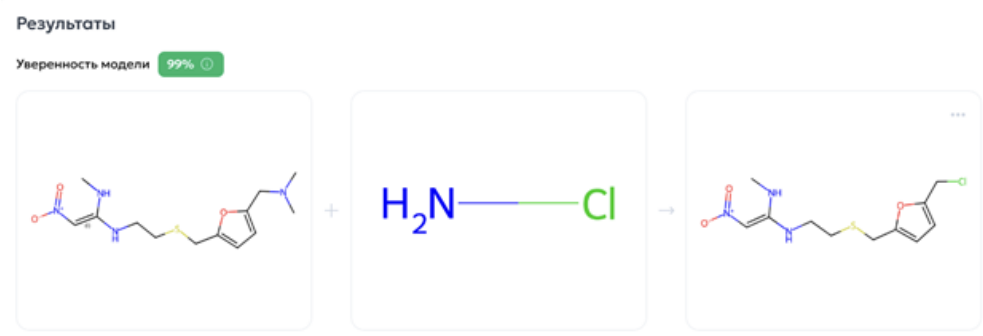


Рис. 47. Побочная реакция взаимодействия ранитидина с хлорамином

В результате побочной реакции образуется диметиламин, который потом под действием HNO_2 окисляется до NDMA. В контексте разложения ранитидина, HNO_2 может образовываться как промежуточный продукт при взаимодействии нитритов с хлорсодержащими окислителями (HOCl, NH_2Cl и др.) (рис 48).

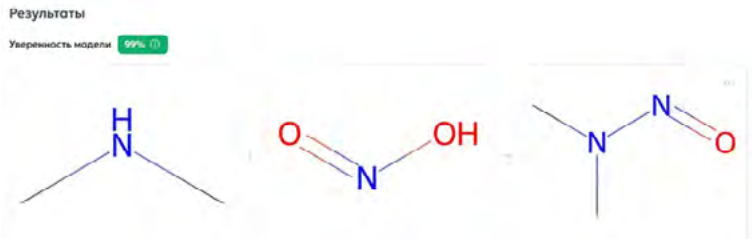


Рис. 48. Образование NDMA

Традиционные методы контроля, такие как лабораторные исследования (ЖХ-МС/МС), требуют значительных затрат времени, ресурсов и финансов, отзыв препаратов происходит постфактум, когда пациенты уже подверглись риску.

Использование хемоинформатики для решения этой проблемы позволяет предсказывать риск образования нитрозаминов на этапе разработки препаратов, минимизируя необходимость в длительных лабораторных тестах.

Метод позволяет за минуты (рис. 49) оценить канцерогенный потенциал побочного продукта CN(C)N=O, а также спрогнозировать показатели его острой и хронической токсичности – задачи, традиционно требующие месяцев экспериментальных исследований.

Рис. 49. Оценка параметров токсичности для N-нитрозодиметиламина (NDMA) с помощью платформы Синтелли [41]

В данном случае: сокращение времени разработки и исключение «проблемных» синтетических путей, снижение затрат (минимизация дорогостоящих лабораторных тестов).

Это один из примеров прогнозирования и предотвращения содержания опасных примесей. Использование хемоинформатики позволяет выявлять такие риски заранее, без длительных лабораторных исследований.

5.4.3. ПФАС – скрытая угроза косметики

Говоря о токсикологии, существует ряд субстанций, распространенность в обществе которых, вызывает особую настороженность у экспертов. Среди них ПФАС (перфторалкильные и полифторалкильные соединения, от англ. PFAS). В марте 2022 года FDA опубликовал любопытную заметку на эту тему [147]. Согласно ней 35 типов ПФАС были обнаружены в 578 косметических средствах. Полный перечень ПФАС можно найти в первоисточнике.

Но почему регулятора беспокоит наличие этих веществ в косметике? Давайте используем инструменты вычислительной токсикологии, чтобы ответить на этот вопрос. В данном случае, спрогнозируем токсикологические параметры вещества с помощью платформы «Синтелли» [41]. Для полноты анализа следовало бы изучить все 35 веществ, однако для иллюстрации принципа нам будет достаточно только одного. Рассмотрим октафторпентилметакрилат, его карточка со свойствами приведена на рисунке 50. И к нашему удивлению, модель прогнозирует репродуктивную токсичность, казалось бы, безобидного вещества. Опытные химики могут возразить, ведь молекула является не рядовым ПФАСом, а акцептором Михаэля. Однако, в действительности, аналогичный прогноз будет для большинства представителей данного класса.

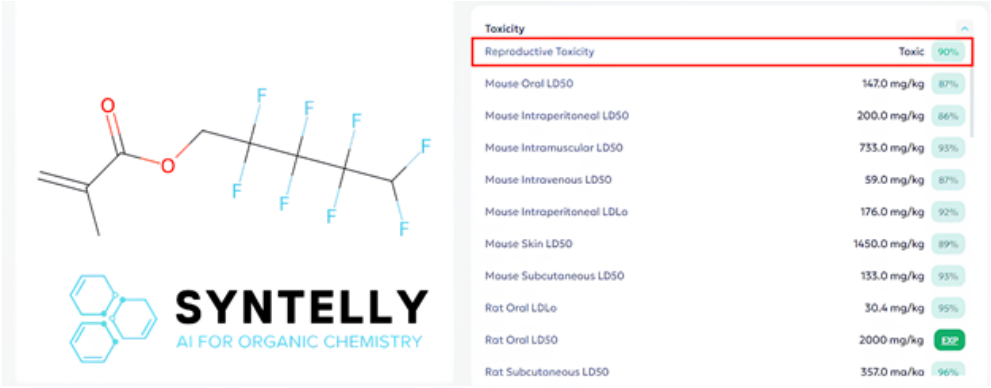


Рис. 50. Предсказанные токсикологические свойства октафторпентилметакрилата с помощью платформы Синтелли [41]

Переставать ли пользоваться косметикой или нет – решать вам. Но правда в том, что зачастую мы даже не осознаем влияние на наше здоровье веществ из продуктов ежедневного пользования.

5.4.4. Грейпфрутовый яд

Довольно известным медицинским фактом является то, что нельзя запивать лекарства грейпфрутовым соком. На эту тему, действительно, существует ряд опубликованных исследований. Кроме того, она получила развитие и так, например, позже были выявлены аналогичные риски от других цитрусовых [148]. Это связано с тем, что в фруктах содержатся фуранокумарины, которые ингибируют ферменты семейства цитохрома P450 (CYP). Иначе говоря, ингибирование этих ферментов приводит к тому, что они медленнее метаболизируют лекарства, что снижает скорость выведения последних. За счет этого концентрация лекарства выше, чем предполагалось. Это, в свою очередь, ведет к нежелательным реакциям организма.

При чем же тут хемоинформатика? Дело в том, что ингибирование CYP ферментов – один из наиболее популярных параметров для прогнозирования. Существует множество инструментов под эту задачу. Благодаря ним Вы можете подтвердить механизм «грейпфрутового ингибирования» в несколько нажатий клавиш. Итак, давайте проверим бергамоттин (представитель фуранокумаринов), который, как считается, и ответственен за описанный эффект. Для этого мы также можем воспользоваться платформой Синтелли [41]. Вводим в соответствующую строку SMILES бергамоттина,

CC(=CCC/C(=C/COC1=C2C=CC(=O)OC2=CC3=C1C=CO3)/C)C,

и в появившейся карточке вещества нас интересует блок «Bio». Вуаля, модель прогнозирует ингибирование бергамоттином 4-х изоформ CYP (рис. 51).

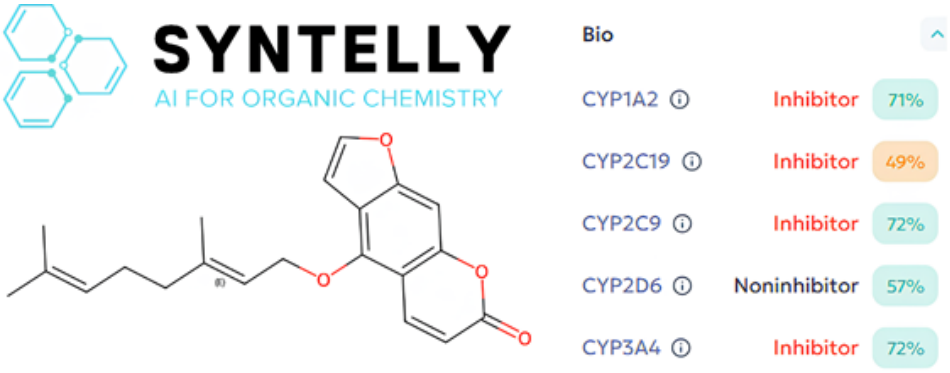


Рис. 51. Прогнозирование ингибирования ферментов CYP бергамоттином с помощью «Синтелли» [41]

5.4.5. Микотоксины

Нами и нашими коллегами проведено исследование плесневого гриба *Albifimbria verrucaria* [149], который является фитопатогеном с широким спектром хозяев, включая важные сельскохозяйственные культуры – овощи

и декоративные растения. Этот гриб продуцирует макроциклические трихотеценовые микотоксины, обладающие высоким токсигенным потенциалом, сопоставимым с афлатоксинами B1 и G. Учитывая их значительную токсичность для сельскохозяйственных животных, как уже упоминалось ранее, возник повышенный интерес к анализу этих микотоксинов и прогнозированию их токсичности.

В таблице 8 представлены микотоксины гриба *Albifimbria verrucaria*, ранжированные по показателю «LD_{50} (мышь, перорально, мг/кг)». Прогнозирование токсикологических параметров в случае отсутствия экспериментальных данных выполнено с использованием платформы на основе искусственного интеллекта Синтелли, которая применяет многозадачные методы машинного обучения [41].

В частности, рассчитаны следующие параметры для оценки токсичности микотоксинов:

- LD_{50} при пероральном введении мышам (RMSE = 0,45 log10(мг/кг))
- LD_{50} при интраперитонеальном введении мышам (RMSE = 0,49 log10(мг/кг))
- LD_{50} у крыс (RMSE = 0,62 log10(мг/кг))
- Период полувыведения у человека (Human Pharmacological Half-life) (ROC AUC = 0,89).

Табл. 9. Микотоксины *Albifimbria verrucaria* (данные взяты из [149])

№ п/п	*Albifimbria verrucaria*	LD_{50} (мышь, перорально, мг/кг)	LD_{50} (мышь, интраперитонеально, мг/кг)	LD_{50} (крыса, перорально, мг/кг)	Период полувыведения у человека	Экспериментальные значения
1	Diacetoxyscirpenol[1]	7,3	7,8	8,34	Низкий / Low	+
2	Roridin A	9	0,5	91,5	Низкий / Low	+
3	Roridin L	29,4	6,93	17,9	Высокий / High	–
4	Verrucarin M	33	8,02	8,3	Высокий / High	–
5	Roridin M	42,3	7,12	12	Высокий / High	+/–*
6	Verrucarin A	43,7	0,5	46,7	Низкий / Low	+
7	Verrucarin B	46,6	13,5	60,3	Низкий / Low	+
8	Roridin K acetate	48,3	11,1	13,6	Высокий / High	–
9	Roridin E Acetate	48,8	13,4	62,8	Высокий / High	–
10	Isororidin-E	52,6	9,67	59,9	Высокий / High	–
11	Roridin E	55	10	59,9	Высокий / High	+
12	Trichoverrin B	61,9	21,8	68,8	Низкий / Low	+/–*
13	Verrucarin J	62,8	7,77	47,7	Высокий / High	+
14	Trichoverrin C	65,4	22,1	63,3	Низкий / Low	-–

Окончание таблицы 9

№ п/п	*Albifimbria verrucaria*	LD_{50} (мышь, перорально, мг/кг)	LD_{50} (мышь, интраперитонеально, мг/кг)	LD_{50} (крыса, перорально, мг/кг)	Период полувыведения у человека	Экспериментальные значения
15	8-Acetylneosolaniol	71,8	34	4,31	Низкий / Low	+
16	Anguidin[2]	109	82,3	8,34	Низкий / Low	+
17	Trichoverrol B	118	92,6	183	Низкий / Low	–
18	Verrucarin E	564	327	2410	Низкий / Low	+/–*
19	7-Hydroxy-3- methoxyviridicatin	1160	337	3700	Низкий / Low	–

Примечание.
*+/- – были проведены экспериментальные исследования по отношению к опухолевым клеткам.
[1] – SMILES: CC(=O)OC[C@]12CCC(C)=C[C@H]1O[C@@H]1[C@H](O)[C@@H](OC(C)=O)[C@@]2 (C)[C@]12CO2.
[2] – SMILES: CC(=O)OC[C@]12CCC(C)=C[C@H]1O[C@@H]1[C@H](O)[C@@H](OC(C)=O)[C@@]2 (C)[C@@]12CO2.

Прогнозируемые данные из таблицы 9 показывают, что около половины микотоксинов, вырабатываемых плесневыми грибами, относятся к I и II классам опасности, что указывает на их высокую токсичность и значительный риск для здоровья человека и животных. Кроме того, большинство микотоксинов этого гриба обладают длительным периодом полувыведения, что свидетельствует о возможности их накопления в организме и продолжительном токсическом эффекте. Без машинного обучения регуляторы не смогут быстро определять приоритетные цели для исследований. Внедрение таких подходов позволит точнее оценивать риски воздействия токсичных веществ на здоровье человека и окружающую среду. Исследование акцентирует внимание на необходимости комплексной оценки токсикологических угроз и дальнейших научных разработок для повышения безопасности. Применение методов хемоинформатики открывает новые возможности для выявления приоритетных соединений, требующих детального изучения экспертами в токсикологии и микологии. Для более глубокого изучения описанного исследования, рекомендуем перейти к прочтению статьи [149].

5.4.6. Прогнозирование токсичности

Исследование, представленное в статье Керстина фон Борриса (Kerstin von Borries) с коллегами [150], посвящено решению важной задачи в оценке химических рисков – отсутствию многих важных данных по токсичности для большинства из более чем 100 000 химических веществ, находящихся в обороте по

всему миру. Для заполнения этого пробела авторы разработали модели машинного обучения (МО) с учетом неопределенности, способные прогнозировать неканцерогенные точки обнаружения (POD) токсичности для более чем 130 000 химических веществ, многие из которых ранее практически не изучались на токсичность. Главная новизна работы заключается в интеграции количественной оценки неопределённости непосредственно в прогнозы моделей: они предоставляют хорошо откалиброванные 95%-ные доверительные интервалы, которые соответствуют фактическим ошибкам. Такой подход повышает прозрачность и способствует формированию доверия к применению машинного обучения в регуляторных и практических задачах, связанных с безопасностью химических веществ.

Особенностью данного исследования является тщательная оценка двух методов машинного обучения с учётом неопределённости: частотного конформного прогнозирования и вероятностных байесовских нейронных сетей. Результаты показали, что модели на основе конформного прогнозирования, особенно с использованием случайных лесов, превосходят байесовские нейронные сети как по точности прогнозов, так и по качеству оценки неопределённости, что подтверждается меньшими ошибками и более надёжными доверительными интервалами. Модели учитывают как алеаторную (связанную с данными), так и эпистемическую (связанную с моделью) неопределённость, что обеспечивает устойчивость прогнозов даже для химических веществ, выходящих за пределы традиционной области применимости стандартных методов машинного обучения. Важным результатом исследования стало не только предоставление первых оценок токсичности для огромного числа химических соединений, но и выявление «горячих точек» с высокой токсичностью и высокой степенью неопределённости прогнозов. Это позволяет определить приоритетные направления для дальнейшего сбора данных и совершенствования моделей. Такие результаты имеют большое значение для эффективного управления рисками и повышения качества и надёжности применения машинного обучения в прогнозировании токсичности химических веществ. Возможно, подобные методы в недалеком будущем будут активно применяться для обновления и составления регистров потенциально опасных химических веществ, таких как Федеральный регистр потенциально опасных химических и биологических веществ [151].

5.4.7. Антидот от перца

Одним из неожиданных открытий 2025 года стало обнаружение потенциальных химических «антидотов» на жгучесть перца чили [152]. Это кажется почти алхимией – найти молекулы, способные укротить огненную силу капсаицина, того самого соединения, что заставляет нас ощущать пламя во рту от халапеньо или каролинского жнеца. И если раньше единственным спасением было молоко или хлеб, то теперь наука предлагает более изящное решение. Исследователь Дэвин Петерсон и его коллеги из Flavor Research and Education Center (Огайо, США) предложил в качестве таких соединений капсианозид I, розеозид и имбирный гликолипид A. В сочетании с капсаицином и дигидро-

капсаицином, которые определяют интенсивность жгучести по шкале Сковилла, эти вещества значительно снижают ощущение жжения.

Данное открытие имеет несколько важных применений. Во-первых, оно открывает возможности для создания новых вкусовых модификаторов, позволяющих регулировать восприятие остроты. Во-вторых, перспективно разработка фармакологических средств, способных десенсибилизировать рецепторы TRPV1 и блокировать болевые ощущения без повреждения тканей. В-третьих, это может способствовать разработке эффективных методов защиты от слезоточивого газа.

Кроме того, с помощью современных платформ, таких как Синтелли [41], возможно моментально предсказать токсичность и безопасность перечисленных антидотов. Это позволяет оценить потенциальные риски и оптимизировать дальнейшие исследования и разработку препаратов на основе этих соединений, что ускоряет процесс их внедрения и минимизирует возможные побочные эффекты. В таблице 9 представлены прогнозируемые показатели токсичности по некоторым параметрам. Представленные данные в таблице 10 демонстрируют низкую токсичность исследованных соединений-антидотов.

Табл. 10. Токсичность соединений-антидотов

Антидот	**LD_{50} (мышь, перорально, мг/кг)**	**LD_{50} (мышь, интраперитонеально, мг/кг)**	**LD_{50} (крыса, внутривенно, мг/кг)**	**Канцерогенность**	**Гепатотоксичность**
Капсианозид I	2780	592	561	нет	нет
Розеозид	1250	459	1210	нет	нет
Имбирный гликолипид A	3970	454	406	нет	нет

5.4.8. Чай с ромашкой или без?

Итак, мы проиллюстрировали ряд способов применить хемоинформатику для познания окружающего мира. Однако этим всё не ограничивается. Ещё одним крайне полезным инструментом является прогнозирование спектра биологической активности соединения. Давайте в качестве примера разберем всеми любимую ромашку. Ее цветки продают в аптеках, она входит в состав различных чаев и продуктов. Считается, что ромашка благоприятно действует на организм. Но как именно? Давайте используем нашим инструменты.

Сперва, нам надо выбрать соединение для скрининга. Известно, что одним из мажорных компонентов является ароматическое соединения хамазулен. В общем-то первая часть названия созвучна с латинским названием ромашки – *Chamomilla*, а вторая с соответствующим ароматическим углеводородом – азуленом. Соответствующий SMILES выглядит так CCC1=CC2=C(C=CC2=C(C=C1)C)C.

Теперь нам нужен инструмент, который предсказывает спектр биологической активности по структуре. И, к нашему счастью, такой существует. Это программа открытого доступа PASS online, разработанная российскими учёными из ИБМХ им. В.Н.Ореховича под руководством академика РАН В.В. Поройкова [153]. С помощью PASS online можно прогнозировать более 4 000 видов биологической активности. Это мощный хемоинформатический инструмент, получивший признание на международной арене.

Итак, вводим SMILES в предназначенное для этого окошко и запускаем расчёт. Спустя мгновение получаем довольно обширный перечень эффектов для хамазулена. Среди них противосудорожное, антисеборейное, противоязвенное действия. Так, мы выяснили, что ромашка может быть полезна и для борьбы с перхотью.

6. БУДУЩЕЕ ХЕМОИНФОРМАТИКИ И ВЫЧИСЛИТЕЛЬНОЙ ТОКСИКОЛОГИИ: ОТКРЫТЫЕ ВОПРОСЫ

6.1 Новые методы анализа рисков химических соединений

Обсуждая вопросы, куда движется такая наука, как химическая информатика в плане использования ее инструментов для задач вычислительной токсикологии, нельзя не затронуть тему Next Generation Risk Assessment (NGRA) - методов нового поколения в оценке рисков, связанных с токсикологическим действием химических веществ на человека и экосистему [154]. На протяжении всей книги мы много говорили о необходимости построения точных моделей для предсказания токсикологических свойств молекул. Обсуждали концепции и подходы того, как это делать. Однако масштаб проблемы наиболее ярко подсветим в этой главе.

Эксперименты на животных до сих пор являются краеугольным камнем споров в научной среде [155]. И каждый на это смотрит по-своему. Зоозащитники говорят о том, что это неэтично. Люди, занимающиеся распределением бюджетов в фармацевтических компаниях, говорят, что это дорого (около 30% всего бюджета тратится на доклинические испытания). Другие специалисты из области скажут, что это долго (доклинические испытания могут занимать от одного года до трех лет). С каждых их этих утверждений нельзя не согласиться. Мы же, в свою очередь, подчеркнем, что данные, полученные при проведении таких экспериментов, часто не являются достоверно надежными.

Человек со средним весом в 75 кг мало напоминает крысу. Согласно некоторым исследованиям, только около 10% результатов доклинических испытаний на животных успешно проходят клинические исследования на людях [156]. Это обусловлено многими факторами. Во-первых, у животных и людей разная скорость и механизмы метаболизма веществ. Например, препарат, безопасный для грызунов, может быть токсичным для человека. Во-вторых, гены, отвечающие за восприимчивость к болезням или реакцию на лечение, часто отличаются. Например, при сходстве генома шимпанзе и человека в 96%, у первых отсутствуют такие болезни, как СПИД, Гепатит В и малярия в принципе. То есть всего 4% различий в общем геноме дают возможность шимпанзе исключить такие социально значимые болезни для человека [157]. Или мыши, с которыми у нас сходство генома не менее 85%, не знают о болезни Альцгеймера или Паркинсона. Для изучения этих болезней у них создают трансгенные модели, которые имитируют патологию (экспрессия белка-предшественника амилоида APP в случае с болезнью Альцгеймера) [158].

В лабораторных экспериментах подавляющее большинство животных составляют мыши и крысы. Эти грызуны широко применяются в биомедицинских исследованиях благодаря своей доступности, высокой плодовитости и генетической однородности. И по результатам исследований, лишь 40%–50% побочных эффектов лекарственных средств у человека можно достоверно воспроизвести на этих лабораторных животных [159]. Это позволяет утверждать, что в некоторых ситуациях случайный выбор с помощью подброшенной монетки был бы даже более информативным. Кроме того, только в 60% случаев мыши и крысы адекватно отражают канцерогенное или эмбриотоксическое воздействие препаратов [160]. Вдогонку можно сказать о том, что как минимум лабораторный стресс, который испытывают животные, находясь на испытаниях, также может повлиять на результаты исследований (доказано, что мужчины-экспериментаторы оказывают больший стресс на животное) [161].

В связи со всеми фактами, изложенными выше, ЕС и США было принято решение о частичном сокращении испытаний на животных в пользу интеграции вычислительных моделей и *in vitro* тестов [162].

В свою очередь российский проект Синтелли, о котором мы уже упоминали, является одним из мировых лидеров в сфере искусственного интеллекта в области органической химии и предлагает пользователям прогноз по нескольким десяткам показателей токсичности для разных организмов. С ее помощью можно сформировать свой собственный датасет из интересующих соединений и предсказать токсикологические характеристики каждой молекулы с видимой точностью прогноза. Такой подход позволяет оценивать первичные показатели токсичности за секунды, что может способствовать сокращению времени и стоимости конечных экспериментов при доклинических испытаниях.

Направление NGRA стремительно развивается, каждый год появляются новые инструменты, концепции и гипотезы, которые в будущем позволят исключить эксперименты на животных и сократить процесс разработки лекарств или принятие решений о жёстком регулировании того или иного потенциально опасного класса веществ [163].

Говоря о тех инструментах, которые в последнее время всё чаще используются исследователями для построения таких моделей, можно отметить графовые нейронные сети (Graph Neural Networks – GNN). На протяжении всей книги мы, по сути, говорили о представлении молекул (дескрипторы, линейные нотации, графы, фингерпринты и т.д.), которое могло бы полностью объяснить их свойства. И вот, появились они – нейросети, которые на вход могут принимать граф с метками на ребрах и вершинах, которые помимо структуры, помогают кодировать физико-химическую информацию отдельных фрагментов (атомов, связей) [164].

Ранние концепции обработки графов с помощью нейросетей начали появляться ещё в конце 1990-ых годов [165]. Однако именно сейчас, с развитием идей об адаптации свёрточных операций для графов и появлением трансформеров, GNN стали более востребованным инструментом. Казалось, что может быть лучше? Подать сам граф молекулы в нейросеть без разбивок ее на фраг-

менты и иметь целостную структуру химического соединения при обучении моделей. Однако исследования, включая работу Ву и его коллег, показали, что в задачах классификации токсичности классические подходы – например, комбинация физико-химических и топологических дескрипторов с традиционными ML-моделями (в частности, SVM) – могут превосходить графовые нейросети по точности [166].

Авторы данной статьи подчеркивают, что такие результаты объяснимы недостаточно большим набором данных токсикологических исследований, который не дает раскрыться всему потенциалу GNN. На текущем этапе при относительно небольших и не очень разнообразных наборах данных по токсичности традиционные методы машинного обучения (SVM, Random Forest) действительно часто показывают более высокие значения AUC, чем глубокое обучение. Преимущество глубоких нейросетей проявляется только при наличии больших объемов данных и сложных задач, где требуется автоматическое извлечение признаков и моделирование сложных зависимостей.

В том числе, одним из наиболее прогрессивных за последнее время направлений в химической информатике является генеративная технология. Использование генеративный сетей (GAN - Generative Adversarial Networks) позволяет создавать молекулы с заданными свойствами [167, 168]. Ярким примером использования такой технологии является ранее упомянутый нами кейс Insilico Medicine с их молекулой INS018_055. В статье [169] впервые раскрывают все этапы разработки от идентификации мишени с помощью ИИ-платформы PandaOmics и генерации структуры молекулы через Chemistry42, до доклинических и клинических испытаний.

Такие тенденции в сторону использования методов глубокого обучения в том числе объяснимы успехом больших языков моделей как ChatGPT и BERT, в основе технологии которых также лежат трансформеры и глубокое обучение [170]. Современные технологии искусственного интеллекта способны существенно ускорить открытие лекарств, снизить затраты и повысить вероятность успеха, открывая новую эру в фармацевтических исследованиях и оценке рисков.

Однако ключевым вопросом на пути дальнейшего развития этого направления является развитие эффективных методов сбора и анализа экспериментальных данных, а также *всесторонний анализ качества* предсказательных моделей, используемых для анализа рисков. Действительно, выше мы несколько раз обсуждали колоссальную асимметрию между относительно низкой скоростью получения экспериментальных данных по ADMET свойствам молекул и скоростью синтеза новых веществ и генерации новых структур с помощью компьютеров. На сегодняшний день этот разрыв достиг трёх порядков: экспериментальные данные о токсичности доступны лишь для одной молекулы из тысячи [149]. Проблема усложняется фрагментарностью баз данных и культурой «закрытости» в ряде компаний (и целых отраслей) относительно доступа к экспериментальным ADMET данным, а также неравномерностью качества используемых предсказательных моделей [171].

Одним из путей частичного решения этой проблемы является организация специализированных хакатонов, которые, как показало недавнее исследование российских учёных и разработчиков [171] позволяют эффективно (то есть быстро и с относительно небольшими затратами) решать сразу несколько задач, связанных с

а) сбором и курированием данных (большое количество команд участников в состоянии «просканировать» большое пространство имеющихся баз данных и литературных источников;

б) объективным и *прозрачным* анализом качества предсказательных моделей (во время хакатона эксперты имеют возможность анализировать все этапы сбора и обработки данных, а также всесторонне валидировать модели);

в) популяризацией области и мотивацией перспективных исследователей к дальнейшей работе в этом направлении.

6.2. Этические аспекты и безопасность данных

Этические вопросы в среде, где информация становится инструментом, который может принести как благо, так и вред, очень важны. На первый взгляд может показаться (особенно специалистам в традиционных технических и естественно-научных областях), что подобные «вечные» вопросы диалектического характера с прицелом на гуманитарные науки сильно субъективно отражают те проблемы, которые на самом деле существуют в этой сфере, а тем более пути их решения (или, попросту, не относятся к сути излагаемых в книге вопросов). Однако не стоит недооценивать идеи, которые в настоящее время предлагают специалисты в области этики использования искусственного интеллекта и безопасности данных. На многие вопросы пока нет однозначного ответа, и именно этим занимаются компетентные специалисты [172]. Что же это за вопросы?

Представим, что искусственный интеллект сгенерировал молекулу, которая, к примеру, лечит от всех видов рака. Кому принадлежат права на эту молекулу? Искусственному интеллекту? Разработчику этой ИИ системы? Или может быть автору исследования, который ввел правильный запрос? Вопрос интеллектуальной собственности на результаты, полученные с помощью искусственного интеллекта, сегодня остается открытым и активно обсуждается как в научном, так и в юридическом сообществе [173]. На практике патентное право большинства стран не признает искусственный интеллект субъектом права, а значит, права на открытие или изобретение чаще всего принадлежат либо разработчику алгоритма, либо организации, которой он принадлежит, либо пользователю, начавшему генерацию. Однако на стыке технологий и права возникают новые вызовы: как учитывать вклад алгоритма, если он работает как «чёрный ящик», и как обеспечить справедливое распределение выгод между всеми участниками процесса?

Кроме вопросов авторства, в хемоинформатике особое значение приобретает проблема ответственности. Действительно, кто должен нести ответственность если сгенерированная модель или молекула приводит к негативным последстви-

ям [174]. Например, к созданию токсичного или просто опасного соединения, кто несет за это ответственность? Разработчик алгоритма, пользователь, предоставивший исходные данные, или, возможно, производитель программного обеспечения? Эти вопросы особенно остро стоят в условиях открытого доступа к инструментам генерации химических структур, где потенциально опасные соединения могут быть созданы и использованы без должного контроля. Примерами таких соединений легко могут стать синтетические наркотики, структуры которых незначительно отличаются друг от друга, но при этом иметь схожие эффекты [175]. Как правило, такие наркотики требуют определенного количества времени со стороны государства для их выявления, идентификации, анализа токсикологических показателей и правового регулирования [176]. Пока все эти стадии будут пройдены, за это время с помощью искусственного интеллекта возможно сгенерировать тысячи таких веществ, отслеживать которые только экспериментальными методами неэффективно с точки затрат временных и финансовых ресурсов. Именно поэтому безопасность данных также становится ключевым аспектом в области искусственного интеллекта в химии.

Хемоинформатика оперирует большими массивами экспериментальных, биологических и фармацевтических данных, которые могут быть чувствительными или коммерчески значимыми [177]. Утечка такой информации может привести к финансовым потерям, нарушению конфиденциальности пациентов или даже к угрозе национальной безопасности, если речь идет о данных, связанных с разработкой новых лекарств или опасных веществ. Поэтому вопросы хранения, передачи и обработки данных должны решаться с учетом современных стандартов кибербезопасности и соответствующего законодательства (например, GDPR в Европе или ФЗ-152 в России). Наконец, нельзя забывать и о социальной ответственности исследователей и разработчиков. В эпоху стремительного развития искусственного интеллекта важно не только следовать букве закона, но и задумываться о возможных последствиях внедрения новых технологий. Это касается как предотвращения злоупотреблений (например, создания «дизайнерских наркотиков», упомянутых ранее, или химического оружия), так и обеспечения равного доступа к достижениям науки для всего общества.

Таким образом, этические аспекты и вопросы безопасности данных в хемоинформатике требуют междисциплинарного подхода, объединяющего усилия химиков, информатиков, юристов, специалистов по кибербезопасности и этике. Только совместная работа позволит выработать эффективные механизмы регулирования и обеспечить ответственное развитие этой важной области науки.

6.3. Платформы на основе машинного обучения и больших данных

В настоящее время всё более заметным трендом в химической информатике является переход от узкоспециализированных инструментов к созданию вычислительных платформ на основе машинного обучения и больших данных (зарубежные Synthia/Chematica [178], Bioptic [179], Pharma.ai [180], отечествен-

ная Синтелли [41]). Ниже мы перечислим отобранные нами ключевые критерии для современного программного обеспечения в химической информатике, подчеркивающие специализированные требования и логически подводящие к преимуществам платформ по отношению к узкоспециализированным инструментам. Стоит отметить, что авторам пока неизвестна платформа, которая бы *полностью* удовлетворяла *всем* перечисленным ниже критериям. Тем не менее, обсуждаемая область обладает высоким потенциалом и быстро развивается, поэтому появление такой платформы возможно в ближайшем будущем.

Основные критерии «идеальной» платформы для химической информатики:

1.Точное представление и манипуляция химическими структурами: поддержка стереохимии, таутомерии, резонансных форм, металлоорганических соединений, полимеров, смесей, биомакромолекул и др.

Важность: без точного представления все последующие расчёты (поиск, прогноз свойств) теряют смысл.

2. Расширенные возможности поиска: не просто поиск по подструктуре, а поиск по *сходству* (разные метрики); поиск по фармакофору; поиск по реакционной способности; поиск различных конформаций; поиск аналогов с учетом «скачков» свойств; поиск как по рецензируемой литературе и патентам, так и по неструктурированным источниками и т.п.

Важность: ключ к открытию новых соединений, анализу экономического и технологического потенциала веществ, а также патентным исследованиям.

3. Расчёт молекулярных дескрипторов и физико-химических свойств: широкий спектр дескрипторов (топологические, геометрические, электронные, поверхностные), точный расчёт физико-химических свойств (logP, pKa, растворимости и др.), конформационного ансамбля, аналитических характеристик веществ (параметры ИК-, ЯМР-спектров и т.д.)

Важность: основа для QSAR/QSPR-моделирования, прогноза ADME/Tox, виртуального скрининга.

4. Интеграция с моделями QSAR/QSPR и Машинного Обучения: возможность строить, валидировать (внутренне/внешне) и применять модели прогноза активности, токсичности и других свойств. Поддержка современных алгоритмов Машинного Обучения (Random Forest, SVM, нейронные сети, в т.ч. GNN).

Важность: предсказательное моделирование - основа современной хемоинформатики для ускорения R&D.

5. Анализ данных и визуализация: мощные инструменты для анализа химического пространства (карты сходства, кластеризация), визуализации свойств, анализа результатов виртуального скрининга, интеграции с «омиксными» данными.

Важность: извлечение знаний из больших химических и биологических данных.

6. Управление химическими данными (Chemical Data Management – CDM [175]): надежное хранение, аннотирование, поиск и извлечение структур, реакций, свойств, биологических данных из больших корпоративных или публичных баз. Поддержка популярных стандартов (SMILES, InChI, RXN и др.).

Важность: без эффективного CDM невозможна работа с большими объемами данных; для многих приложений (особенно в фарме), надёжность и безопасность данных являются очень *чувствительными* показателями.

7. Прогнозирование реакций и синтеза с соответствующей специализацией: предсказание продуктов реакции, выходов, ретросинтетический анализ, оценка синтетической доступности (SA), планирование синтетических путей.

Важность: критично для дизайна синтеза новых соединений и оптимизации процессов.

Традиционное ПО часто фокусируется на отдельных аспектах (структуры, поиск, QSAR, CDM). Однако главный вызов современной химии - интеллектуальное проектирование молекул с заданными свойствами и эффективное планирование их синтеза. Это требует интеграции всех перечисленных выше возможностей в единую сквозную платформу, обогащенную искусственным интеллектом и экспертными знаниями.

Таким образом, платформы типа Синтелли предоставляют пользователям целый ряд новых возможностей:

1. Глубокая интеграция знаний: они кодируют огромные массивы *экспертных знаний* о химических реакциях (правила, ограничения, условия, выходы) в машинно-читаемый формат, дополняя это данными из литературы и патентов.

2. Мощный ИИ для ретросинтеза: используют сложные алгоритмы (часто на основе графовых нейронных сетей и символьного ИИ) для *предсказания осуществимых и эффективных синтетических путей*, учитывая не только химическую логику, но и доступность реагентов, стоимость, безопасность, время.

3. Сквозной рабочий процесс: позволяют от *дизайна целевой молекулы* (с учетом прогноза свойств/токсичности) через *ретросинтетический анализ* перейти к *планированию конкретных экспериментов* в одной вычислительной среде.

4. Обучаемость и адаптивность: способны обучаться на новых данных (успешных/неуспешных синтезах, токсикологических экспериментах и пр.), постоянно улучшая свои предсказания.

5. Фокус на практической реализуемости: критерии выбора путей синтеза включают не только соответствие законам химии, но и *синтетическую доступность (SA), стоимость, время, безопасность, экологичность* процесса.

6. Ускорение и оптимизация R&D: кардинально сокращают время на планирование синтеза, позволяют находить более короткие/дешевые/безопасные пути, избегать тупиковых направлений.

Вывод:

Хотя традиционные инструменты хемоинформатики (для управления данными, поиска, расчёта дескрипторов, QSAR) остаются важными «рабочими лошадками», в будущем мы ожидаем рост интереса к интеллектуальным интегрированным платформами типа Синтелли. Они преодолевают разрозненность специализированных инструментов, объединяя глубокие химические знания, мощь ИИ и фокус на практической реализуемости для решения ключевой задачи: не просто предсказать, *что* можно синтезировать, а определить, *как* это сделать оптимальным способом. Это превращает хемоинформатику из инструмента анализа в инструмент *интеллектуального дизайна и принятия решений*, что является качественным скачком в развитии химической науки и промышленности, а также химической безопасности.

6.4. Исследования экспосома и химическая информатика

В заключении этой главы, мы хотели бы немного поговорить о понятии экспосома и его роли в современных масштабных токсикологических исследованиях с помощью инструментов химической информатики. Экспосом –динамическая совокупность всех экзогенных факторов (химические, биологические, физические), воздействующих на организм в течение жизни [181, 182]. Это понятие включает загрязнители воздуха, пестициды, бытовую химию, микробиомные компоненты и другие агенты, взаимодействующие с геномом и всем организмом в целом и влияющие на здоровье. В отличие от относительно статичного генома, экспосом вариабелен во времени и пространстве, что делает его изучение критически важным для понимания причин хронических заболеваний, таких как рак, диабет или нейродегенеративные расстройства, а также целый ряд хронических эндокринных нарушений.

Экспосом объясняет «неизвестные» этиологические факторы болезней. Например, исследования NIEHS (США) выявили, что комбинации промышленных химикатов даже при низких дозах вызывают эндокринные нарушения, ведущие к репродуктивным расстройствам и онкологии [183].

Такие проекты, как *Human Exposome Project* [184] интегрируют данные экспосома с эпидемиологией, создавая карты экологических рисков для уязвимых групп (дети, беременные, пожилые люди). Хемоинформатика в данном контексте предоставляет методы для систематизации и прогнозирования взаимодействий между компонентами экспосома и биологическими мишенями:

А. Управление данными и классификация

- Базы химических структур и анализ структурных классов: специализированные базы данных (см. выше) хранят аннотированные данные о структурах, свойствах и токсичности соединений. Далее хемоинформатические инструменты могут быть использованы для категоризации вещества по механизмам действия (например, эстрогенная активность) на основе дескрипторов [185]. Для маркеров экспосома (например, бисфенолы, фталаты) можно создать

«структурные правила» токсичности. Таким образом, можно идентифицировать молекулы-эндокринные разрушители по наличию фенольных групп или галогенированных фрагментов.

Б. Прогнозирование токсичности

- QSAR-модели: предсказывают активность соединений по дескрипторам [186].
- Докинг и молекулярное моделирование: для изучения механизмов действия. В работе [183] докинг показал, что тетрахлорэтилен связывается с лиганд-связывающим доменом ERα, объясняя его эстрогенную активность при концентрациях ниже порога токсичности.

В. Интеграция с «омиксными» данными

Хемоинформатика позволяет соединять химические данные с биологическими системными моделями на разных уровнях описания живых организмов и соответствующих метаболических путей. Мы приводим внизу всего лишь пару примеров, но исследования в этом направлении ведутся сейчас очень активно по всему миру:

- многомасштабные сетевые сценарии взаимодействий молекул в живых системах: *In silico* реконструкция путей токсичности [187].
- Мультиомный анализ: совместное моделирование экспосома и транскриптома/метаболома [185].

ЗАКЛЮЧЕНИЕ

Подводя итоги данного пособия, хочется верить, что оно помогло Вам погрузиться в удивительный мир химической информатики со всеми ее противоречиями и достижениями. Эта область действительна достойна того, чтобы о ней писали, говорили и обсуждали по всему миру. Во времена, когда цифровизация проникает во все сферы нашей жизни, нельзя игнорировать новые инструменты, которые помогают ускорять процесс разработки новых лекарств, материалов и технологий, сохранять природу и позволяют сделать наш мир более безопасным. Хемоинформатика уже сегодня меняет облик современной науки, открывая двери к тем решениям, которые ещё недавно казались невозможными в плане быстрого и точного анализа химических рисков и разработки принципиально новых веществ и материалов.

Однако путь в этой области не всегда будет легким. Вам наверняка встретятся сложные задачи, неоднозначные результаты, а иногда и разочарования. Но именно в такие моменты рождаются настоящие открытия и формируется подлинное мастерство. Не бойтесь задавать вопросы, экспериментировать, ошибаться и искать нестандартные подходы. Наука любит смелых и любознательных.

Помните, что хемоинформатика – это не только про алгоритмы и базы данных, но и про людей, идеи и сотрудничество. Особенно важно наладить сотрудничество между экспериментальными группами (данные!), прикладными математиками и программистами (алгоритмы и программы), собственно, хемоинформатиками (всё из вышеперечисленного + построение предсказательных

моделей и их валидация), а также представителями компаний и регуляторов (постановка задач и выделение ресурсов на их решение). Здесь всегда есть место для творчества, для поиска новых смыслов и большого количества инноваций. Ваша настойчивость, интерес и желание учиться – вот главные инструменты, которые помогут вам двигаться вперед.

Пусть это пособие станет для Вас не только источником знаний, но и отправной точкой для собственных оригинальных исследований, открытий и изобретений. Мир химической информатики открыт для новых идей и ждет именно вашего вклада.

Междисциплинарная команда ИППИ РАН им. А. А. Харкевича желает Вам вдохновения, профессионального роста и удовольствия от каждого шага на этом захватывающем пути!

БЛАГОДАРНОСТИ

Авторы выражают искреннюю признательность многим коллегам и экспертам, чьи ценные замечания и профессиональные рекомендации значительно способствовали улучшению содержания этой книги. Особая благодарность Алиеву Т.А., Бодрову А.Д., Гуниной П.В., Изотовой М.Е., Игнатьеву А.А., Кислинскому А.С., Климовой А.С., Кормановской Е.Б., Матюхину А.Ю., Мухамеджановой А.А., Надеину А.В., Осолодкину Д.И., Петрову Ф.И., Пинигиной А.Е., Поздееву А.В., Поройкову В.В., Сафронову В.С., Скорб Е.В., Соснину С.Б., Стешину И.С., Туманову А.В., Федоровой В.В., Шкилю Д.О., Шульге Д.А.

ГЛОССАРИЙ

1. ADMET

Аббревиатура, описывающая ключевые свойства лекарственных молекул: абсорбция (Absorption), распределение (Distribution), метаболизм (Metabolism), выведение (Excretion), токсичность (Toxicity).

2. AlphaFold

Алгоритм ИИ для предсказания трехмерной структуры белков на основе аминокислотной последовательности.

3. Applicability Domain (Домен применимости)

Область химического пространства, в которой модель QSAR/QSPR даёт достоверные прогнозы.

4. Биологическая активность

Способность вещества вызывать биологический ответ. В работах по медицинской химии этим термином часто обозначают концентрацию вещества, требуемую для достижения специфического эффекта заданной интенсивности по сравнению с определенным стандартом.

5. CAS (Chemical Registry System) – номер
Уникальный числовой идентификатор, присваиваемый каждому химическому веществу в базе данных Chemical Abstracts Service для однозначной идентификации.

6. CDM (Chemical Data Management) – управление химическими данными
Термин включает в себя надежное хранение, аннотирование, поиск и извлечение структур, реакций, свойств, биологических данных из больших корпоративных или публичных баз. Поддержка популярных стандартов и форматов данных.

7. Дескрипторы молекулы
Числовые характеристики, описывающие физико-химические или структурные свойства молекулы (например, молекулярная масса, logP).

8. Докинг (молекулярный докинг)
Компьютерное моделирование взаимодействия малой молекулы (лиганда) с биологической мишенью (например, белком) для предсказания их связывания.

9. Fingerprints (Фингерпринты)
Битовая строка, отражающая наличие или отсутствие определенных структурных фрагментов в молекуле (MACCS Keys, ECFP и т.д.).

10. Gradient Boosting (Градиентный бустинг)
Метод машинного обучения, где модели строятся последовательно для исправления ошибок предыдущих.

11. GUI (Graphical User Interface)
Графический пользовательский интерфейс, представляющий собой способ взаимодействия пользователя с компьютером или программой через визуальные элементы, такие как окна, кнопки, меню, иконки и другие графические компоненты.

12. Гидрофобность/Липофильность
Свойство молекулы избегать контакта с водой, характеризуемое коэффициентом распределения октанол-вода (logP).

13. In silico
Эксперимент или анализ, проводимый с помощью компьютерного моделирования, а не в лаборатории (in vitro) или на живых организмах (in vivo).

14. InChI (International Chemical Identifier)
Уникальный идентификатор химических соединений, позволяющий стандартизировать их представление в цифровом формате.

15. Кросс-валидация
Метод оценки качества модели машинного обучения, который заключается в многократном разбиении исходного набора данных на обучающие и тестовые части с последующим обучением и проверкой модели на разных подвыборках.

16. Лиганд
Молекула, которая связывается с молекулярной мишенью и вызывает, блокирует или изменяет определенный биологический ответ.

17. QSAR (Quantitative Structure-Activity Relationship)
Метод, устанавливающий количественную связь между структурой молекулы и её биологической активностью.

18. QSPR (Quantitative Structure-Property Relationship)
Аналогично QSAR, но для прогнозирования физико-химических свойств молекул.

19. Random Forest (Случайный лес)
Ансамблевый метод машинного обучения, использующий множество деревьев решений для классификации или регрессии.

20. RDKit
Библиотека с открытым исходным кодом для обработки химических данных и машинного обучения.

21. REACH (Registration Evaluation Authorisation Restriction – Регистрация Оценка Разрешение Ограничение)
Регламент ЕС по регистрации и оценке безопасности химических веществ.

22. SMILES (Simplified Molecular Input Line Entry System)
Линейная нотация для однозначного описания структуры молекулы в виде строки символов.

23. Скрининг
Массовое тестирование или анализ большого числа соединений для выявления среди них активных или обладающих нужными свойствами.

24. Structural Alerts (Структурные признаки токсичности)
Фрагменты молекулы, связанные с повышенной токсичностью или другими нежелательными эффектами.

25. Токсикологические эндпоинты
Конкретные виды токсических эффектов (например, канцерогенность, нейротоксичность).

26. Вычислительная токсикология

Наука о применении математических и компьютерных моделей для прогнозирования неблагоприятных последствий и лучшего понимания отдельных или множественных механизмов, посредством которых данное химическое вещество причиняет вред.

27. Хемоинформатика (Химическая информатика)

Мультидисциплинарная область, объединяющая химию, информатику и математику для анализа и обработки химических данных, включая прогнозирование свойств молекул и разработку лекарств.

СПИСОК ЛИТЕРАТУРЫ

1. Hann M., Green R. Chemoinformaticsa new name for an old problem? // Current Opinion in Chemical Biology. – 1999. – Т. 3. – №. 4. – С. 379–383.

2. Bajorath J. (ed.). Chemoinformatics: concepts, methods, and tools for drug discovery. – Springer Science & Business Media, 2008. – Т. 275.

3. Brown F. K. et al. Chemoinformatics: what is it and how does it impact drug discovery //Annual reports in medicinal chemistry. – 1998. – Т. 33. – С. 375–384.

4. Gasteiger J. Handbook of chemoinformatics : from data to knowledge in 4 volumes. Handbook of chemoinformatics. – Wiley-VCH, 2003.

5. Voigt K., Welzl G. Chemical databases: an overview of selected databases and evaluation methods //Online Information Review. – 2002. – Т. 26. – №. 3. – С. 172–192.

6. Willett P. Chemoinformatics: a history //Wiley Interdisciplinary Reviews: Computational Molecular Science. – 2011. – Т. 1. – №. 1. – С. 46–56.

7. Tate F. A. Chemical abstracts service //American Journal of Health-System Pharmacy. – 1966. – Т. 23. – №. 2. – С. 63–67.

8. Dearden J. C. The history and development of quantitative structure-activity relationships (QSARs) //Oncology: breakthroughs in research and practice. – IGI Global, 2017. – С. 67–117.

9. Medford A. J. et al. Extracting knowledge from data through catalysis informatics //Acs Catalysis. – 2018. – Т. 8. – №. 8. – С. 7403–7429.

10. Idakwo G. et al. A review on machine learning methods for in silico toxicity prediction //Journal of Environmental Science and Health, Part C. – 2018. – Т. 36. – №. 4. – С. 169–191.

11. Martinez-Mayorga K. et al. The impact of chemoinformatics on drug discovery in the pharmaceutical industry //Expert opinion on drug discovery. – 2020. – Т. 15. – №. 3. – С. 293–306.

12. Reymond J. L. The chemical space project //Accounts of chemical research. – 2015. – Т. 48. – №. 3. – С. 722–730.

13. Subbarayan P. R. Impact and Challenges of Chemoinformatics in Drug Discovery //Artificial Neural Network for Drug Design, Delivery and Disposition. – Academic Press, 2016. – С. 141–152.

14. Jumper J. et al. Highly accurate protein structure prediction with AlphaFold // nature. – 2021. – Т. 596. – №. 7873. – С. 583–589.

15. Meng X. Y. et al. Molecular docking: a powerful approach for structure-based drug discovery //Current computer-aided drug design. – 2011. – Т. 7. – №. 2. – С. 146–157.

16. Engel T. Basic overview of chemoinformatics //Journal of chemical information and modeling. – 2006. – Т. 46. – №. 6. – С. 2267–2277.

17. Lesk A. M. Introduction to bioinformatics. – Oxford university press, 2019.

18. Cramer C. J. Essentials of computational chemistry: theories and models. – John Wiley & Sons, 2013.

19. Parr R. G. Density functional theory of atoms and molecules //Horizons of Quantum Chemistry: Proceedings of the Third International Congress of Quantum Chemistry Held at Kyoto, Japan, October 29-November 3, 1979. – Dordrecht : Springer Netherlands, 1989. – С. 5–15.

20. Wiswesser W. J. A line-formula chemical notation. – Crowell, 1954.

21. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules //Journal of chemical information and computer sciences. – 1988. – Т. 28. – №. 1. – С. 31–36.

22. Heller S. R. et al. InChI, the IUPAC international chemical identifier //Journal of cheminformatics. – 2015. – Т. 7. – С. 1–34.

23. Нейн Ю. И., Иванцова М. Н. Компьютерное представление химической информации: учебное пособие. – 2020.

24. Weininger D., Weininger A., Weininger J. L. SMILES. 2. Algorithm for generation of unique SMILES notation //Journal of chemical information and computer sciences. – 1989. – Т. 29. – №. 2. – С. 97–101.

25. Маджидов Т.И., Баскин И.И., Антипин И.С., Варнек А.А. Введение в хемоинформатику: Компьютерное представление химических структур: учебное пособие. – 2013. – Введение в хемоинформатику.

26. Буркатовская Ю. Б. Теория графов //Издательство Томского политехнического университета. – 2014.

27. Rosa A. et al. On certain valuations of the vertices of a graph //Theory of Graphs (Internat. Symposium, Rome. – 1966. – С. 349–355.

28. Ullmann J. R. Bit-vector algorithms for binary constraint satisfaction and subgraph isomorphism //Journal of Experimental Algorithmics (JEA). – 2011. – Т. 15. – С. 1.1–1.64.

29. Dalby A. et al. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited //Journal of chemical information and computer sciences. – 1992. – Т. 32. – №. 3. – С. 244–255.

30. Bernstein F. C. et al. The Protein Data Bank: a computer-based archival file for macromolecular structures //Journal of molecular biology. – 1977. – Т. 112. – №. 3. – С. 535–542.

31. O'Boyle N. M. et al. Open Babel: An open chemical toolbox //Journal of cheminformatics. – 2011. – Т. 3. – С. 1–14.

32. Kim S. et al. PubChem in 2021: new data content and improved web interfaces //Nucleic acids research. – 2021. – Т. 49. – №. D1. – С. D1388–D1395.

33. Zdrazil B. et al. The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods //Nucleic acids research. – 2024. – Т. 52. – №. D1. – С. D1180–D1192.

34. Баскин И.И., Маджидов Т.И., Варнек А.А. Введение в хемоинформатику: Химические базы данных: учебное пособие. Ч. 2. – Казань: Изд-во Казан. ун-та, 2015. – 188 с.

35. Kim S. et al. PubChem substance and compound databases //Nucleic acids research. – 2016. – Т. 44. – №. D1. – С. D1202–D1213.

36. Wishart D. S. et al. DrugBank 5.0: a major update to the DrugBank database for 2018 //Nucleic acids research. – 2018. – Т. 46. – №. D1. – С. D1074-D1082.

37. Bleicher K. H. et al. Hit and lead generation: beyond high-throughput screening //Nature reviews Drug discovery. – 2003. – Т. 2. – №. 5. – С. 369–378.

38. Irwin J. J., Shoichet B. K. ZINC− a free database of commercially available compounds for virtual screening //Journal of chemical information and modeling. – 2005. – Т. 45. – №. 1. – С. 177–182.

39. Burley S. K. et al. Protein Data Bank (PDB): the single global macromolecular structure archive //Protein crystallography: methods and protocols. – 2017. – С. 627–641.

40. Gaulton A. et al. The ChEMBL database in 2017 //Nucleic acids research. – 2017. – Т. 45. – №. D1. – С. D945–D954.

41. Syntelly [Электронный ресурс]. – Режим доступа: https://syntelly.ru (дата обращения: 27.05.2025).

42. Wu L. et al. TOXRIC: a comprehensive database of toxicological data and benchmarks //Nucleic acids research. – 2023. – Т. 51. – №. D1. – С. D1432–D1445.

43. Якубке Х. Д., Ешкайт Х. Аминокислоты, пептиды, белки. – М.: Мир, 1985. – Т. 457.

44. Финкельштейн А. В., Птицын О. Б. Физика белка //М.: Книжный дом «Университет. – 2002. – Т. 41.

45. Овчинников Ю. А. Биоорганическая химия. – Рипол Классик, 1987.

46. DeLano W. L. et al. Pymol: An open-source molecular graphics tool //CCP4 Newsl. Protein Crystallogr. – 2002. – Т. 40. – №. 1. – С. 82–92.

47. Theobald O. Machine learning for absolute beginners: a plain English introduction. – Scatterplot press, 2021.

48. Звягинцева А.В., Павленко А.А. Основы токсикологии: учебное пособие / А.В. Звягинцева, А.А. Павленко. — Воронеж: ФГБОУ ВПО «Воронежский государственный технический университет», 2012. – 251 с.

49. Sajid N. A. et al. Single vs. multi-label: The issues, challenges and insights of contemporary classification schemes //Applied Sciences. – 2023. – Т. 13. – №. 11. – С. 6804.

50. Wolf, Andrew. Machine Learning Simplified: A Gentle Introduction to Supervised Learning. 2022.

51. Никлаус В. Алгоритмы и структуры данных. – Litres, 2022.

52. Миронов А. М. Машинное обучение. Часть 1 //М.: МАКС Пресс. – 2018. – С. 4.

53. Efron B. Bootstrap methods: another look at the jackknife //Breakthroughs in statistics: Methodology and distribution. – New York, NY : Springer New York, 1992. – С. 569–593.

54. Breiman L. Bagging predictors //Machine learning. – 1996. – Т. 24. – С. 123–140.

55. Баскин И. И., Маджидов Т. И., Варнек А. А. Введение в хемоинформатику: учеб. пособие. Ч. 4. Методы машинного обучения //Казань: Изд-во Казан. ун-та. – 2016.

56. Friedman J. H. Greedy function approximation: a gradient boosting machine //Annals of statistics. – 2001. – С. 1189–1232.

57. ChemDraw [Электронный ресурс]. URL: https://perkinelmer-chemdraw-professional.software.informer.com/ (дата запроса: 28.05.2025)

58. ChemAxon Ltd. Chemaxon [программное обеспечение]. URL: https://chemaxon.com (дата обращения: 28.05.2025).

59. Landrum G. Rdkit documentation //Release. – 2013. – Т. 1. – №. 1–79. – С. 4.

60. The RDKit Documentation – The RDKit 2025.03.1 documentation. — URL: https://www.rdkit.org/docs/index.html (дата обращения: 10.06.2025).

61. Bento A. P. et al. An open source chemical structure curation pipeline using RDKit //Journal of Cheminformatics. – 2020. – Т. 12. – С. 1–16.

62. Cherkasov A. et al. QSAR modeling: where have you been? Where are you going to? //Journal of medicinal chemistry. – 2014. – Т. 57. – №. 12. – С. 4977–5010.

63. Höskuldsson A. PLS regression methods //Journal of chemometrics. – 1988. – Т. 2. – №. 3. – С. 211–228.

64. Cortes C., Vapnik V. Support-vector networks //Machine learning. – 1995. – Т. 20. – С. 273–297.

65. Lo Y. C. et al. Machine learning in chemoinformatics and drug discovery // Drug discovery today. – 2018. – Т. 23. – №. 8. – С. 1538–1546.

66. Stone M. Cross□validatory choice and assessment of statistical predictions // Journal of the royal statistical society: Series B (Methodological). – 1974. – Т. 36. – №. 2. – С. 111–133.

67. Ferreira L. L. G., Andricopulo A. D. ADMET modeling approaches in drug discovery //Drug discovery today. – 2019. – Т. 24. – №. 5. – С. 1157–1165.

68. DiMasi J. A., Grabowski H. G., Hansen R. W. Innovation in the pharmaceutical industry: new estimates of R&D costs //Journal of health economics. – 2016. – Т. 47. – С. 20–33.

69. Rang H.P. Rang and Dale's pharmacology. — Edinburgh ; New York : Elsevier/Churchill Livingstone, 2012. – 806 с.

70. Bender A., Cortés-Ciriano I. Artificial intelligence in drug discovery: what is realistic, what are illusions? Part 1: Ways to make an impact, and why we are not there yet //Drug discovery today. – 2021. – Т. 26. – №. 2. – С. 511–524.

71. Фридрихсберг Д. А. Курс коллоидной химии. – Рипол Классик, 1984.

72. Zhang S. et al. Enteric and hydrophilic polymers enhance dissolution and absorption of poorly soluble acidic drugs based on micro-environmental pH-modifying solid dispersion //European Journal of Pharmaceutical Sciences. – 2022. – Т. 168. – С. 106074.

73. Menichetti R., Kanekal K. H., Bereau T. Drug–membrane permeability across chemical space //ACS central science. – 2019. – Т. 5. – №. 2. – С. 290–298.

74 Llinàs A., Glen R. C., Goodman J. M. Solubility challenge: can you predict solubilities of 32 molecules using a database of 100 reliable measurements? //Journal of chemical information and modeling. – 2008. – Т. 48. – №. 7. – С. 1289–1303.

75. Palmer D. S. et al. First-principles calculation of the intrinsic aqueous solubility of crystalline druglike molecules //Journal of chemical theory and computation. – 2012. – Т. 8. – №. 9. – С. 3322–3337.

76. Li A. P. Accurate prediction of human drug toxicity: a major challenge in drug development //Chemico-biological interactions. – 2004. – Т. 150. – №. 1. – С. 3–7.

77. Temple A. R. Pathophysiology of aspirin overdosage toxicity, with implications for management //Pediatrics. – 1978. – Т. 62. – №. 5s. – С. 873–876.

78. Johnson E. A. Clostridial toxins as therapeutic agents: benefits of nature's most toxic proteins //Annual Reviews in Microbiology. – 1999. – Т. 53. – №. 1. – С. 551–575.

79. Pirazzini M. et al. Botulinum neurotoxins: biology, pharmacology, and toxicology //Pharmacological reviews. – 2017. – Т. 69. – №. 2. – С. 200–235.

80. Raies A. B., Bajic V. B. In silico toxicology: computational methods for the prediction of chemical toxicity //Wiley Interdisciplinary Reviews: Computational Molecular Science. – 2016. – Т. 6. – №. 2. – С. 147–172.

81. Куценко, С. А. (2004). Основы токсикологии

82. Кукин П. П., Пономарев Н. Л., Таранцева К. Р. Основы токсикологии. Учебное пособие. – 2012.

83. Kavlock R. J. et al. Computational toxicology–a state of the science mini review //Toxicological sciences. – 2008. – Т. 103. – №. 1. – С. 14–27.

84. Kavlock R., Dix D. Computational toxicology as implemented by the US EPA: providing high throughput decision support tools for screening and assessing chemical exposure, hazard and risk //Journal of Toxicology and Environmental Health, Part B. – 2010. – Т. 13. – №. 2-4. – С. 197–217.

85. Durant J. L. et al. Reoptimization of MDL keys for use in drug discovery // Journal of chemical information and computer sciences. – 2002. – Т. 42. – №. 6. – С. 1273–1280.

86. Sosnin S. et al. Comparative study of multitask toxicity modeling on a broad chemical space //Journal of chemical information and modeling. – 2018. – Т. 59. – №. 3. – С. 1062–1072.

87. Kalgutkar A. S. Designing around structural alerts in drug discovery //Journal of Medicinal Chemistry. – 2019. – Т. 63. – №. 12. – С. 6276–6302.

88. Judson R. et al. ACToR–aggregated computational toxicology resource // Toxicology and applied pharmacology. – 2008. – Т. 233. – №. 1. – С. 7–13.

89. Williams E. S., Panko J., Paustenbach D. J. The European Union's REACH regulation: a review of its history and requirements //Critical reviews in toxicology. – 2009. – Т. 39. – №. 7. – С. 553–575.

90. Schneider N. et al. Development of a novel fingerprint for chemical reactions and its application to large-scale reaction classification and similarity //Journal of chemical information and modeling. – 2015. – Т. 55. – №. 1. – С. 39–53.

91. Drwal M. N. et al. Molecular similarity-based predictions of the Tox21 screening outcome //Frontiers in Environmental science. – 2015. – Т. 3. – С. 54.

92. Cereto-Massagué A. et al. Molecular fingerprint similarity search in virtual screening //Methods. – 2015. – Т. 71. – С. 58–63.

93. Martin Y. C., Kofron J. L., Traphagen L. M. Do structurally similar molecules have similar biological activity? //Journal of medicinal chemistry. – 2002. – Т. 45. – №. 19. – С. 4350–4358.

94. Steinbeck C. et al. The Chemistry Development Kit (CDK): An open-source Java library for chemo-and bioinformatics //Journal of chemical information and computer sciences. – 2003. – Т. 43. – №. 2. – С. 493–500.

95. Rogers D., Hahn M. Extended-connectivity fingerprints //Journal of chemical information and modeling. – 2010. – Т. 50. – №. 5. – С. 742–754.

96. Bondi A. van der Waals Volumes and Radii //The Journal of physical chemistry. – 1964. – Т. 68. – №. 3. – С. 441–451.

97. Babine R. E., Bender S. L. Molecular recognition of protein− ligand complexes: Applications to drug design //Chemical reviews. – 1997. – Т. 97. – №. 5. – С. 1359–1472.

98. Kier L. B., Hall L. H. An electrotopological-state index for atoms in molecules //Pharmaceutical research. – 1990. – Т. 7. – С. 801–807.

99. Roy K., Kar S., Ambure P. On a simple approach for determining applicability domain of QSAR models //Chemometrics and Intelligent Laboratory Systems. – 2015. – Т. 145. – С. 22–29.

100. Pearson K. LIII. On lines and planes of closest fit to systems of points in space //The London, Edinburgh, and Dublin philosophical magazine and journal of science. – 1901. – Т. 2. – №. 11. – С. 559–572.

101. Bajusz D., Rácz A., Héberger K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? //Journal of cheminformatics. – 2015. – Т. 7. – С. 1–13.

102. Moriwaki H. et al. Mordred: a molecular descriptor calculator //Journal of cheminformatics. – 2018. – Т. 10. – №. 1. – С. 4.

103. Shen M. et al. Application of predictive QSAR models to database mining: identification and experimental validation of novel anticonvulsant compounds // Journal of medicinal chemistry. – 2004. – Т. 47. – №. 9. – С. 2356–2364.

104. Belsley D. A., Kuh E., Welsch R. E. Regression diagnostics: Identifying influential data and sources of collinearity. – John Wiley & Sons, 2005.

105. Weisberg S. Applied linear regression. – John Wiley & Sons, 2005. – Т. 528.

106. Воронина В. В. и др. Теория и практика машинного обучения. – 2017.

107. Cabello-Solorzano K. et al. The impact of data normalization on the accuracy of machine learning algorithms: a comparative analysis //International conference on soft computing models in industrial and environmental applications. – Cham : Springer Nature Switzerland, 2023. – С. 344–353.

108. Starovoitov V. V., Golub Y. I. Data normalization in machine learning // Informatics. – 2021. – Т. 18. – №. 3. – С. 83–96.

109. Mahesh B. et al. Machine learning algorithms-a review //International Journal of Science and Research (IJSR).[Internet]. – 2020. – Т. 9. – №. 1. – С. 381–386.

110. Tropsha A. Best practices for QSAR model development, validation, and exploitation //Molecular informatics. – 2010. – Т. 29. – №. 6–7. – С. 476–488.

111. Raju N. S. et al. Methodology review: Estimation of population validity and cross-validity, and the use of equal weights in prediction //Applied Psychological Measurement. – 1997. – Т. 21. – №. 4. – С. 291–305.

112. Баскин И. И., Маджидов Т. И., Варнек А. А. Введение в хемоинформатику: учеб. пособие. Ч. 3. Моделирование «структура–свойство» //Казань: Изд-во Казан. ун-та. – 2015.

113. Achary P. G. R. Applications of quantitative structure-activity relationships (QSAR) based virtual screening in drug design: a review //Mini Reviews in Medicinal Chemistry. – 2020. – Т. 20. – №. 14. – С. 1375–1388.

114. Broach J. R. et al. High-throughput screening for drug discovery //Nature. – 1996. – Т. 384. – №. 6604. – С. 14–16.

115. Yang S. Y. Pharmacophore modeling and applications in drug discovery: challenges and recent advances //Drug discovery today. – 2010. – Т. 15. – №. 11–12. – С. 444–450.

116. Schaller D. et al. Next generation 3D pharmacophore modeling //Wiley Interdisciplinary Reviews: Computational Molecular Science. – 2020. – Т. 10. – №. 4. – С. e1468.

117. Sanders M. P. A. et al. From the protein's perspective: the benefits and challenges of protein structure-based pharmacophore modeling //MedChemComm. – 2012. – Т. 3. – №. 1. – С. 28–38.

118. Bender A., Glen R. C. Molecular similarity: a key technique in molecular informatics //Organic & biomolecular chemistry. – 2004. – Т. 2. – №. 22. – С. 3204–3218.

119. Miteva M. A. et al. FAF-Drugs: free ADME/tox filtering of compound collections //Nucleic acids research. – 2006. – Т. 34. – №. suppl_2. – С. W738–W744.

120. Nicolaou K. C. et al. The art and science of total synthesis at the dawn of the twenty□first century //Angewandte Chemie International Edition. – 2000. – Т. 39. – №. 1. – С. 44–122.

121. Прокудина Н. В., Крамчанинов М. М. Клинические случаи эффективного применения агностического подхода в терапии опухолей с мутацией BRAF V600E //Злокачественные опухоли. – 2024. – Т. 14. – №. 1. – С. 92–98.

122. Menzies A. M., Long G. V., Murali R. Dabrafenib and its potential for the treatment of metastatic melanoma //Drug design, development and therapy. – 2012. – С. 391–405.

123. Wang Y. et al. Retrosynthesis prediction with an interpretable deep-learning framework based on molecular assembly tasks //Nature Communications. – 2023. – Т. 14. – №. 1. – С. 6155.

124. Dong J. et al. Deep learning in retrosynthesis planning: datasets, models and tools //Briefings in Bioinformatics. – 2022. – Т. 23. – №. 1. – С. bbab391.

125. Herbst R. S., Fukuoka M., Baselga J. Gefitinib–a novel targeted approach to treating cancer //Nature Reviews Cancer. – 2004. – Т. 4. – №. 12. – С. 956–965.

126. Баскин И.И., Маджидов Т.И., Варнек А.А. Введение в хемоинформатику: учебное пособие. Часть 5. Информатика химических реакций. – 2017. – Введение в хемоинформатику.

127. Mandal S. Mee'nal Moudgil, and Sanat K. Mandal,» Rational drug design // European journal of pharmacology. – 2009. – Т. 625. – №. 1–3. – С. 90–100.

128. Morris G. M., Lim-Wilby M. Molecular docking //Molecular modeling of proteins. – 2008. – С. 365–382.

129. Lagunin A. et al. QSAR modelling of rat acute toxicity on the basis of PASS prediction //Molecular informatics. – 2011. – Т. 30. – №. 2–3. – С. 241–250.

130. Capdeville R. et al. Glivec (STI571, imatinib), a rationally developed, targeted anticancer drug //Nature reviews Drug discovery. – 2002. – Т. 1. – №. 7. – С. 493–502.

131. Clozel M. et al. Pharmacological characterization of bosentan, a new potent orally active nonpeptide endothelin receptor antagonist //The Journal of pharmacology and experimental therapeutics. – 1994. – Т. 270. – №. 1. – С. 228–235.

132. Gallwitz B. Clinical use of DPP-4 inhibitors //Frontiers in endocrinology. – 2019. – Т. 10. – С. 389.

133. Xu Z. et al. INS018-055, A Novel Traf2-and NCK-interacting Kinase (TNIK) Inhibitor, Improves Lung Function in Patients With Idiopathic Pulmonary Fibrosis: Results From a Randomized, Double-blind, Placebo-controlled Phase 2a Study //American Journal of Respiratory and Critical Care Medicine. – 2025. – Т. 211. – №. Abstracts. – С. A2904–A2904.

134. Devillers J. Methods for building QSARs //Computational Toxicology: Volume II. – Totowa, NJ : Humana Press, 2012. – С. 3–27.

135. Lei P. et al. Green solvent selection for Suzuki–Miyaura coupling of amides //ACS Sustainable Chemistry & Engineering. – 2020. – Т. 9. – №. 1. – С. 552–559.

136. Sunderland E. M. et al. A review of the pathways of human exposure to poly-and perfluoroalkyl substances (PFASs) and present understanding of health effects //Journal of exposure science & environmental epidemiology. – 2019. – Т. 29. – №. 2. – С. 131–147.

137. Ciallella H. L., Zhu H. Advancing computational toxicology in the big data era by artificial intelligence: data-driven and mechanism-driven modeling for chemical toxicity //Chemical research in toxicology. – 2019. – Т. 32. – №. 4. – С. 536–547.

138. Martinez-Mayorga K., Medina-Franco J. L. Chemoinformatics—applications in food chemistry //Advances in food and nutrition research. – 2009. – Т. 58. – С. 33–56.

139. Dagan-Wiener A. et al. BitterDB: taste ligands and receptors database in 2019 //Nucleic Acids Research. – 2019. – Т. 47. – №. D1. – С. D1179–D1185.

140. Ahmed J. et al. SuperSweet—a resource on natural and artificial sweetening agents //Nucleic acids research. – 2010. – Т. 39. – №. suppl_1. – С. D377–D382.

141. Goel M. et al. FlavorDB2: an updated database of flavor molecules //Journal of Food Science. – 2024. – Т. 89. – №. 11. – С. 7076–7082.

142. Medina-Franco J. L. et al. Chemoinformatic analysis of GRAS (Generally Recognized as Safe) flavor chemicals and natural products //PLoS One. – 2012. – Т. 7. – №. 11. – С. e50798.

143. Garg N. et al. FlavorDB: a database of flavor molecules //Nucleic acids research. – 2018. – Т. 46. – №. D1. – С. D1210–D1216.

144. Sorokina M. et al. COCONUT online: collection of open natural products database //Journal of Cheminformatics. – 2021. – Т. 13. – №. 1. – С. 2.

145. Маджидов Т.И., Баскин И.И., Варнек А.А. Введение в хемоинформатику: Химическое пространство и виртуальный скрининг: учебное пособие. Ч. 6. — Казань: Изд-во Казанского университета, 2019. – 240 с.

146. Shaik K. M. et al. Regulatory updates and analytical methodologies for nitrosamine impurities detection in sartans, ranitidine, nizatidine, and metformin along with sample preparation techniques //Critical reviews in analytical chemistry. – 2022. – Т. 52. – №. 1. – С. 53–71.

147. Commissioner O. of the Per and Polyfluoroalkyl Substances (PFAS) in Cosmetics // FDA. – 2024.

148. Saito M. et al. Undesirable effects of citrus juice on the pharmacokinetics of drugs: focus on recent studies //Drug Safety. – 2005. – Т. 28. – С. 677–694.

149. Ткаченко В.Т., Федоров М.В., Федорова В.В., Поздеев А.В., Кормановская Е.Б., Климова А.С., Гунина П.В. Новые методы оценки рисков патогенов: машинное обучение в анализе спектра токсичности Albifimbria verrucaria // Вестник войск РХБ защиты. – 2025. – Vol. 9. – Новые методы оценки рисков патогенов. – No. 1. – P. 57–73.

150. a) von Borries K. et al. Uncertainty-aware machine learning to predict non-cancer human toxicity for the global chemicals market. – 2025.; b) von Borries K. et al. Potential for Machine Learning to Address Data Gaps in Human Toxicity and Ecotoxicity Characterization// Environmental Science & Technology – 2023. – Т. 57. – №. 46. – С. 18259–18270

151. Федеральный регистр потенциально опасных химических и биологических веществ | НИАЦ РПОХБВ ФБУН «ФНЦГ им. Ф.Ф.Эрисмана» Роспотребнадзора. – URL: https://www.rpohv.ru/online/ (дата обращения: 10.06.2025).

152. Borcherding J., Tello E., Peterson D. G. Identification of Chili Pepper Compounds That Suppress Pungency Perception //Journal of Agricultural and Food Chemistry. – 2025.

153. Way2Drug – main. – URL: https://www.way2drug.com/passonline/ (дата обращения: 10.06.2025).

154. Yang C. et al. The role of a molecular informatics platform to support next generation risk assessment //Computational Toxicology. – 2023. – Т. 26. – С. 100272.

155. Kiani A. K. et al. Ethical considerations regarding animal experimentation //Journal of preventive medicine and hygiene. – 2022. – Т. 63. – №. 2 Suppl 3. – С. E255.

156. Pound P., Bracken M. B. Is animal research sufficiently evidence based to be a cornerstone of biomedical research? //Bmj. – 2014. – Т. 348.

157. Varki A., Altheide T. K. Comparing the human and chimpanzee genomes: searching for needles in a haystack //Genome research. – 2005. – Т. 15. – №. 12. – С. 1746–1758.

158. Sasaguri H. et al. APP mouse models for Alzheimer's disease preclinical studies //The EMBO journal. – 2017. – Т. 36. – №. 17. – С. 2473–2487.

159. Pound P., Ritskes-Hoitinga M. Is it possible to overcome issues of external validity in preclinical animal research? Why most animal models are bound to fail // Journal of translational medicine. – 2018. – Т. 16. – С. 1–8.

160. Olson H. et al. Concordance of the toxicity of pharmaceuticals in humans and in animals //Regulatory toxicology and pharmacology. – 2000. – Т. 32. – №. 1. – С. 56–67

161. Sorge R. E. et al. Olfactory exposure to males, including men, causes stress and related analgesia in rodents //Nature methods. – 2014. – Т. 11. – №. 6. – С. 629–632.

162. Punt A. et al. New approach methodologies (NAMs) for human-relevant biokinetics predictions: Meeting the paradigm shift in toxicology towards an animal-free chemical risk assessment //Altex. – 2020. – Т. 37. – №. 4. – С. 607–622.

163. Pistollato F. et al. Current EU regulatory requirements for the assessment of chemicals and cosmetic products: challenges and opportunities for introducing new approach methodologies //Archives of toxicology. – 2021. – Т. 95. – С. 1867–1897.

164. Boiko D. A. et al. Advancing molecular machine learning representations with stereoelectronics-infused molecular graphs //Nature Machine Intelligence. – 2025. – Т. 7. – №. 5. – С. 771–781.

165. Sperduti A., Starita A. Supervised neural networks for the classification of structures //IEEE transactions on neural networks. – 1997. – Т. 8. – №. 3. – С. 714–735.

166. Wu Y., Wang G. Machine learning based toxicity prediction: from chemical structural description to transcriptome analysis //International journal of molecular sciences. – 2018. – Т. 19. – №. 8. – С. 2358.

167. Goodfellow I. et al. Generative adversarial networks //Communications of the ACM. – 2020. – Т. 63. – №. 11. – С. 139–144.

168. Creswell A. et al. Generative adversarial networks: An overview //IEEE signal processing magazine. – 2018. – Т. 35. – №. 1. – С. 53–65.

169. Ren F. et al. A small-molecule TNIK inhibitor targets fibrosis in preclinical and clinical models //Nature Biotechnology. – 2025. – Т. 43. – №. 1. – С. 63–75.

170. Schwaller P. et al. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction //ACS central science. – 2019. – Т. 5. – №. 9. – С. 1572–1583.

171. Shkil D. O. et al. Expanding Predictive Capacities in Toxicology: Insights from Hackathon-Enhanced Data and Model Aggregation //Molecules. – 2024. – Т. 29. – №. 8. – С. 1826.

172. а) Рекомендация об этических аспектах искусственного интеллекта // ЮНЕСКО – 2021. б) Кожевников Н. Н., Данилова В. С. Этические аспекты искусственного интеллекта //Наука и техника в Якутии. – 2020. – №. 2 (39). – С. 28–31.

173. Понкин И., Редькина А. Искусственный интеллект и право интеллектуальной собственности //Интеллектуальная собственность. Авторское право и смежные права. – 2018. – №. 2. – С. 35–44.

174. Urbina F. et al. Dual use of artificial-intelligence-powered drug discovery // Nature machine intelligence. – 2022. – Т. 4. – №. 3. – С. 189–191.

175. Luethi D., Liechti M. E. Designer drugs: mechanism of action and adverse effects //Archives of toxicology. – 2020. – Т. 94. – №. 4. – С. 1085–1133.

176. Sacco L. N., Finklea K. M. Synthetic drugs: overview and issues for congress. – Congressional Research Service, Library of Congress, 2011.

177. Khan N. A. et al. Cyber Security and Privacy Safeguarding Pharmaceutical Innovation in a Digital Age: Pharmaceutical Innovation in a Digital Age //Pakistan BioMedical Journal. – 2025. – С. 02–10.

178. Discovery at Your Fingertips | SYNTHIA® Retrosynthesis Software. — URL: https://www.synthiaonline.com/ (дата обращения: 10.06.2025).

179. BiOptic Inc. – URL: https://www.bioptic.com.tw/ (дата обращения: 10.06.2025).

180. PHARMA.AI. – URL: https://pharma.ai (дата обращения: 10.06.2025).

181. Wild C. P. Complementing the genome with an “exposome”: the outstanding challenge of environmental exposure measurement in molecular epidemiology //Cancer Epidemiology Biomarkers & Prevention. – 2005. – Т. 14. – №. 8. – С. 1847–1850.

182. Wild C. P. The exposome at twenty: a personal account //Exposome. – 2025. – Т. 5. – №. 1. – С. osaf003.

183. Alofe O. et al. Determining the endocrine disruption potential of industrial chemicals using an integrative approach: Public databases, in vitro exposure, and modeling receptor interactions //Environment international. – 2019. – Т. 131. – С. 104969.

184. Hartung T. A call for a human exposome project //ALTEX-Alternatives to animal experimentation. – 2023. – Т. 40. – №. 1. – С. 4–33.

185. Schneider M. et al. In silico predictions of endocrine disruptors properties // Endocrinology. – 2019. – Т. 160. – №. 11. – С. 2709–2716.

186. Rahu I., Kull M., Kruve A. Predicting the activity of unidentified chemicals in complementary bioassays from the HRMS data to pinpoint potential endocrine disruptors //Journal of Chemical Information and Modeling. – 2024. – Т. 64. – №. 8. – С. 3093–3104.

187. La Merrill M. A. et al. Consensus on the key characteristics of endocrine-disrupting chemicals as a basis for hazard identification //Nature Reviews Endocrinology. – 2020. – Т. 16. – №. 1. – С. 45–57.

И.А. Моргунов, И.Д. Никитин, В.Т. Ткаченко, М.В. Федоров

ХЕМОИНФОРМАТИКА И ВЫЧИСЛИТЕЛЬНАЯ ТОКСИКОЛОГИЯ

Формат 70x100 1/16
Гарнитура Times
Усл.-п. л. 10,4. Уч.-изд. л. 9,5
Тираж 300 экз.

Издатель – Российская академия наук

Публикуется в авторской редакции

Верстка и печать – УНИД РАН
Отпечатано в экспериментальной цифровой типографии РАН

Издается по решению Научно-издательского совета
Российской академии наук (НИСО РАН) от 13.02.2025 № 19
и распространяется бесплатно