МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
имени М. В. ЛОМОНОСОВА
ХИМИЧЕСКИЙ ФАКУЛЬТЕТ

На правах рукописи

Загрибельный Богдан

Платформа генеративной химии в моделировании структур потенциальных лекарственных веществ

1.4.16. Медицинская химия
1.4.3. Органическая химия

ДИССЕРТАЦИЯ
на соискание ученой степени
кандидата химических наук

Научные руководители:
кандидат химических наук,
старший научный сотрудник
Палюлин Владимир Александрович

кандидат биологических наук
Иваненков Ян Андреевич

Москва – 2025

Оглавление

Введение

Актуальность и степень разработанности темы исследования. Поиск новых малых лекарственных молекул остается краеугольным камнем разработки потенциальных лекарственных веществ. Малые лекарственные молекулы, как правило, характеризующиеся низкой молекулярной массой, обладают уникальной способностью взаимодействовать с биологическими макромолекулами, такими как белки, ДНК и даже РНК [1], модулируя их функцию таким образом, что это может привести к терапевтическим эффектам. Эта универсальность делает малые молекулы незаменимыми в лечении широкого спектра заболеваний, от инфекционных болезней и рака до неврологических и аутоиммунных расстройств.

Развитие технологий искусственного интеллекта (ИИ) и машинного обучения (МО) значительно ускорило поиск новых малых лекарственных молекул, несмотря на то, что пока созданные при помощи ИИ потенциальные лекарственные вещества не были зарегистрированы национальными профильными регуляторами [2]. Тем не менее, эти технологии позволяют быстро анализировать огромные наборы данных, прогнозировать свойства и оптимизировать процесс разработки препаратов, снижая затраты и время, связанные с разработкой лекарств. В то же время, создание новых лекарственных веществ всё ещё является крайне трудоемкой и дорогой задачей, требующей многопараметрической оптимизации, которая помимо чисто химических и фармакологических требований к фармсубстанции, включает и факторы иной природы, такие как экономическую целесообразность (баланс между размером рынка и затратами на разработку и производство с учетом возможной конкуренции) и юридические аспекты в лице особенностей патентной конкуренции и регистрации препаратов. Интеграция учета всех факторов, прямо или косвенно влияющих на принятие решение в рамках разработки потенциальных лекарственных веществ, в рамках платформенных решений на основе искусственного интеллекта представляется для профессионального сообщества ключом к проблемам индустрии. Описанию одной из первых в мире таких платформ и созданию её ключевых узлов, связанных с моделированием структурных трендов медицинской химии и оценкой синтезируемости молекулярных структур посвящена настоящая диссертация.

Разработанная платформа генеративной химии Chemistry42 является первым в своем роде инструментом решения задач генеративной химии. По этой причине профессиональное сообщество не было консолидировано относительно того, какие сценарии моделирования структур потенциальных лекарственных веществ могут выполняться при помощи подобной платформы.

Предлагаемое в диссертации теоретическое определение понятия о синтетической доступности на языке теории вероятности ранее не было описано в литературе. Разработанный метод моделирования синтетической доступности молекулярных структур ReRSA (*Retrosynthesis-Related Synthetic Accessibility*, англ., синтетическая доступность, связанная с ретросинтезом) является первым описанным методом, учитывающим одновременно ретросинтетический, статистический и дескрипторный факторы, сочетает в себе удовлетворительную точность и высокую скорость, делающую его первым подобным методом, в контексте задач генеративной химии.

Ранее в литературе не были описаны молекулярные дескрипторы, способные давать оценку соответствия молекулярных структур трендам, наблюдаемым в медицинской химии. Необходимость балансировать структурную новизну в терминах соответствия текущему состоянию развития медицинской химии и синтетическую доступность генерируемых молекулярных структур требовало создания вышеупомянутого молекулярного дескриптора.

Цель работы заключается в обеспечении разработанной платформы генеративной химии надежными алгоритмами оценки синтезируемости и соответствия структурным трендам медицинской химии, а также модельными экспериментами, иллюстрирующими функциональность платформы с позиции базовых сценариев моделирования структур потенциальных лекарственных веществ.

Для достижения указанной цели были поставлены следующие **задачи**:

1.Добиться того, чтобы модельные эксперименты покрывали большую часть базовых сценариев компьютеризированного дизайна малых лекарственных молекул и позволили пользователям эффективно овладеть функционалом платформы генеративной химии.

2.Учесть в рамках разработки нового метода моделирования синтетической доступности лучшие стороны существующих подходов и сделать для нового метода удобную визуализацию в рамках пользовательского интерфейса в целях повышения интерпретируемости результатов.

3.Создать набор данных из молекулярных структур, запатентованных крупнейшими фармацевтическими компаниями, с учётом хронологического порядка в целях более точного моделирования структурных трендов и соответствующую функцию, описывающие эти тренды.

Объектами исследования являлись молекулярные структуры зарегистрированных и потенциальных лекарственных веществ — малых лекарственных молекул.

Предметом исследования являлось моделирование синтетической доступности молекулярных структур зарегистрированных и потенциальных лекарственных веществ, моделирование структурных трендов, наблюдаемых в медицинской химии, а также

моделирование молекулярных структур в соответствии с базовыми сценариями компьютеризированного дизайна малых лекарственных молекул.

Методология и методы исследования. Создание и первичная валидация программного кода MCE-18 (*Medicinal Chemistry Evolution*, англ., эволюция медицинской химии) и ReRSA выполнялось на основе хемоинформатической библиотеки RDKit на языке программирования Python. Вторичная валидация методов осуществлялась на платформе генеративной химии Chemistry42 в рамках модельных экспериментов.

Научная новизна. В настоящей диссертации впервые предложен метод моделирования синтетической доступности, объединяющий элементы статистического анализа встречаемости фрагментов в референсном химическом пространстве, ретросинтетический анализ и дескрипторный подход к моделированию структурной сложности. Часть функционала метода, анализирующая встречаемость 5-членных ароматических гетероциклов, использует новаторские хемоинформатические идеи, в частности, автоматизированную генерацию больших библиотек SMARTS-подструктур, компрессию SMARTS-строк по атомным примитивам, иерархический поиск по SMARTS-подструктурам.

При разработке молекулярного дескриптора MCE-18 было впервые обозначено различие при анализе sp^3-гибридизированных атомов углерода на предмет их включенности в кольцевые системы. Данное различие вошло в основу дескриптора NCSPTR, компонента дескриптора MCE-18, который выгодно отличается от классического дескриптора Fsp^3, который, в свою очередь, ассоциируется с успехом в клинических испытаниях. В отличие от Fsp^3, дескриптор MCE-18 эффективно описывает структурную эволюцию химического пространства потенциальных лекарственных веществ.

Теоретическая и практическая ценность работы. Разработан теоретический аппарат на основе теории вероятности для сферы знаний о синтетической доступности. Метод моделирования синтетической доступности ReRSA запатентован и интегрирован в платформу генеративной химии Chemistry42, которой пользуются крупнейшие фармацевтические компании мира, включая Roche, Merck, Elly Lilly, BMS, Arvinas, UCB, Takeda и другие.

В рамках дизайна и валидации платформы генеративной химии Chemistry42 были созданы модельные эксперименты, иллюстрирующие базовые сценарии ранней разработки потенциальных лекарственных веществ. Данные модельные эксперименты могут быть взяты за основу для разработки и валидации любой другой платформы генеративной химии или быть использованы для создания бенчмаркинг-платформы, оценивающей эффективность генеративных моделей. Метод оценки соответствия структурным трендам медицинской химии MCE-18 интегрированный в платформу позволяет пользователям понять, насколько

генерируемые молекулярные структуры соответствуют современным трендам в разработке потенциальных лекарственных веществ.

Положения, выносимые на защиту.

1. Модельные эксперименты, созданные в целях иллюстрации функционала разработанной платформы генеративной химии Chemistry42 в рамках базовых сценариев компьютеризированного дизайна малых лекарственных молекул, позволяют проводить обучение на платформе и её валидацию.

2. Разработанный метод моделирования синтетической доступности ReRSA позволяет эффективно генерировать на платформе молекулярные структуры с высокой ожидаемой синтетической осуществимостью.

3. Концепция полноподструктурного анализа молекулярных структур, которая была разработана в ходе создания метода моделирования синтетической доступности ReRSA и провалидирована на примере пятичленных ароматических гетероциклов, обладает большим потенциалом для хемоинформатической области и может быть распространена на любой другой класс подструктур (циклы, линкеры, периферические фрагменты).

4.Разработанный молекулярный дескриптор MCE-18 позволяет создавать на платформе генеративной химии Chemistry42 молекулярные структуры, соответствующие трендам современной медицинской химии, которые задают крупнейшие фармацевтические компании.

Достоверность полученных результатов обеспечивается использованием для расчетов стандартных, широко используемых, статистически обоснованных алгоритмов и программного обеспечения, публикациями в рецензируемых научных изданиях.

Личный вклад автора состоит в подборе, анализе и систематизации литературы, постановке промежуточных задач. Автор принимал непосредственное участие в создании важнейших элементов платформы генеративной химии, включая модули оценки синтезируемости молекулярных структур и оценки соответствия трендам современной медицинской химии, в обработке и интерпретации экспериментального материала, подготовке материалов к публикации в научных журналах, написании патентных заявок. Во всех опубликованных в соавторстве работах по теме диссертационной работы вклад автора (Загрибельного Б.) является основополагающим, в том числе в работе [1], где автором проделана работа сбору обучающей выборки ингибиторов DDR1 киназы для генеративной

[1] Zhavoronkov A., Ivanenkov Y.A., Aliper A., Veselov M.S., Aladinskiy V.A., Aladinskaya A.V., Terentiev V.A., Polykovskiy D.A., Kuznetsov M.D., Asadulaev A., Volkov Y., Zholus A., Shayakhmetov R.R., Zhebrak A., Minaeva L.I., **Zagribelnyy B.**, Lee L.H., Soll R., Madge D., Xing L., Guo T., Aspuru-Guzik A. Deep learning enables rapid identification of potent DDR1 kinase inhibitors // *Nature Biotechnology* — 2019. — V. 37. — *No. 9.* — pp. 1038-1040, — DOI: 10.1038/s41587-019-0224-x. — EDN YKXOEF. Импакт-фактор 41.7 (JIF), 0.35 п.л., доля вклада 10%

модели GENTRL; в статье [2] и патенте [3], где автор проделал работу по моделированию структур-кандидатов потенциальных ингибиторов главной протеазы SARS-CoV-2 на платформе Chemistry42, отбору структур на синтез, подготовке патентной заявки, оптимизации соединений-хитов и, в целом, по руководству всем проектом по разработке упомянутых ингибиторов, начиная с идентификации соединений-хитов, вплоть до номинирования лидирующей серии соединений; в патенте [4], где автор проделал работу по химической концептуализации движка для автоматизированного ретросинтеза молекулярных структур.

Апробация работы и публикации. По результатам работы опубликованы 6 статей в рецензируемых научных журналах, индексируемых в базе ядра РИНЦ «eLibrary Science Index», международными базами данных (Web of Science, Scopus, RSCI) и рекомендованных для защиты в диссертационном совете МГУ для публикации результатов диссертационных работ по специальностям 1.4.16. Медицинская химия (химические науки) и 1.4.3. Органическая химия (химические науки) и 3 патента. Результаты, полученные в ходе проделанной работы, были представлены на XI Международной конференции молодых ученых по химии “Mendeleev–2019”, 9–13 сентября 2019. г. Петергоф. По результатам конференции автор настоящей диссертации был удостоен третьей премии “за лучший устный доклад”. Также по материалам диссертации был представлен доклад на IV Международном форуме об искусственном интеллекте, робототехнике, инновациях в образовании и подготовке кадров “Digital Innopolis Days 2024”, 2–4 октября 2024. Республика Татарстан, г. Иннополис.

Структура и объем работы. Работа состоит из введения, трех глав, включающих обзор литературы, материалы и методы, результаты и их обсуждение, а также заключения, списка литературы и приложения. Общее количество страниц: 220, включая приложение. Основная часть работы содержит 77 иллюстраций (69 рисунков и 8 синтетических схем) и 41 таблицу; список литературы включает 196 наименований. Приложение, данное на 2 страницах, содержит 1 таблицу.

---

[2] Sun J., Sun D., Yang Q., Wang D., Peng J., Guo H., Ding X., Chen Zh., Yuan B., Ivanenkov Y.A., Yuan J., **Zagribelnyy B.,** He Y., Su J., Wang L., Tang J., Li Zh., Li R., Li T., Hu X., Liang X., Zhu A., Wei P., Fan Y., Liu S., Zheng J., Guan X., Aliper A., Yang M., Bezrukov D.S., Xie Zh., Terentiev V.A., Peng G., Polykovskiy D.A., Malyshev A.S., Malkov M.N., Zhu Q., Aspuru-Guzik A., Ding X., Cai X., Zhang Man, Zhao J., Zhong N., Ren F., Chen X., Zhavoronkov A., Zhao J. A novel, covalent broad-spectrum inhibitor targeting human coronavirus Mpro // *Nature Communications* — 2025. — V. 16. — *P.* 4546, – DOI 10.1038/s41467-025-59870-4. – EDN DTDLMX. Импакт-фактор 15.7 (JIF), 1.15 п.л., доля вклада 25%

[3] Патент № US20230174488A1. Sars-cov-2 inhibitors having covalent modifications for treating coronavirus infections: опубл. 08.06.2023 / Zhavoronkovs A., Ivanenkov Y.A., **Zagribelnyy B.,** 9.03 п.л., доля вклада 40%

[4] Патент № US20220172802A1. Retrosynthesis systems and methods: опубл. 02.06.2022 / Konstantinov A., Putin E.O., **Zagribelnyy B.,** Ivanenkov Y.A., Zhavoronkovs A., 1.96 п.л., доля вклада 30%

Обозначения и сокращения

ВМБ — взаимодействия между белками
ВПС — высокопроизводительный скрининг
ГО — глубокое обучение
ИИ — искусственный интеллект
КДИС — коммерчески доступные исходные соединения
КССА — количественная связь структура–активность
МД — молекулярный дескриптор
ММСД — метод моделирования синтетической доступности
МО — машинное обучение
ПФ — привилегированные фрагменты
СД — синтетическая доступность
СО — синтетическая осуществимость
СОК — самоорганизующиеся карты

ADMET — *absorption, distribution, metabolism, and excretion–toxicity*, абсорбция, распределение, метаболизм, выведение, токсичнсоть
AIDD — *artificial intelligence-assisted drug design*, дизайн потенциальных лекарственных веществ при помощи искусственного интеллекта
CADD — *computer-assisted drug design*, компьютеризированный дизайн потенциальных лекарственных веществ
CAS # — уникальный численный идентификатор химических соединений, внесённый в реестр Chemical Abstracts Service
CASP — *computer-aided synthesis planning*, планирование синтеза с помощью компьютера
CRO — *contract research organization,* контрактная исследовательская организация
DSTA — *design, synthesis, testing and analysis,* дизайт, синтез, тестрирование и анализ
ES — *easy-to-synthesize*, легко синтезируемое
FBDD — *fragment-based drug design*, дизайн потенциальных лекарственных веществ на основе знаний о связывании фрагмента с молекулярной мишенью
FDA — Food and Drug Administration, Управление по контролю качества пищевых продуктов и лекарственных средств США
FGI — *functional group interconversion*, преобразование функциональной группы
HS — *hard-to-synthesize*, трудно синтезируемое
LBDD — *ligand-based drug design*, дизайн потенциальных лекарственных веществ, основанный на знании о структуре лигандов
MCF — *medicinal chemistry filters*, медхимические фильтры
MW — *molecular weight*, молекулярный вес
PDB — Protein Data Bank
Ph4 — *pharmacophore*, фармакофор
PLI — *pocket-ligand interactions*, взаимодействия между лигандом и карманом

PROTAC — *proteolysis targeting chimera*, химерные (комбинированные) соединения, индуцирующие протеолиз мишени

SBDD — *structure-based drug design*, дизайн потенциальных лекарственных веществ, основанный на знании о структуре мишени

1. Обзор литературы

1.1 Проблематика и практики моделирования структур потенциальных лекарственных веществ

1.1.1 Проблематика размерности химического пространства в дизайне молекулярных структур

Понятие “химическое пространство” относится к обширному, практически бесконечному множеству всех возможных химических соединений, охватывающему каждую мыслимую комбинацию атомов и молекулярных структур. Одной из центральных проблем медицинской химии является огромный размер этого химического пространства. По оценкам, количество потенциальных молекул, подобных лекарственным, превышает 10^{60}, что настолько велико, что затмевает общее количество молекул, которые когда-либо могли бы быть синтезированы и протестированы традиционными экспериментальными методами [3]. Этот огромный размер создает значительное препятствие в выявлении новых терапевтических агентов, так как практически невозможно исчерпывающе исследовать это пространство.

Огромные масштабы химического пространства создают серьезные проблемы для разработки потенциальных лекарственных веществ. Во-первых, поиск новых эффективных лекарственных молекул становится подобным поиску “иголки в стоге сена”. Даже при использовании технологий высокопроизводительного скрининга, которые могут оценивать миллионы соединений, доля химического пространства, которую можно практически исследовать, крайне мала. Это ограничение означает, что многие потенциально ценные соединения остаются неоткрытыми, просто потому что методы, используемые для исследования химического пространства, недостаточно всеобъемлющие. Во-вторых, с проблемой размера химического пространства напрямую связана проблема его неоднородности и наличия “разрывов” во многих областях свойств объектов [4]. Даже если объекты находятся близко в химических пространстве в рамках некоторого стандартного представления и метрик, экспериментальные свойства объектов могут отличаться радикально [5].

Кроме того, размер химического пространства поднимает вопрос об относительной предвзятости в выборе хемотипов исследуемых соединений. Усилия по открытию лекарств часто сосредотачиваются на хорошо известных классах молекул, что приводит к исследованию относительно небольшого и часто химически схожего подмножества химического пространства. Такое внимание может ограничивать инновации и снижать шансы

на обнаружение новых соединений с уникальными механизмами действия. В результате существует растущая потребность в стратегиях, которые могут эффективно навигировать по химическому пространству, выделяя области, которые с наибольшей вероятностью содержат перспективных кандидатов, и избегая тех, которые уже переизучены или вряд ли приведут к новым открытиям.

В качестве наглядной иллюстрации несопоставимых размеров всего химического пространства и области химического пространства, соответствующей зарегистрированным лекарственным веществам, можно привести тот факт, что в последней доступной версии на 2023 примерного перечня важнейших лекарственных препаратов Всемирной Организации Здравоохранения содержится 591 наименование лекарств и 103 терапевтических аналогов [6]. Перечень обновляется каждые два года и прирост числа новых наименований в каждой новой версии не превышает в среднем 30 единиц, причем не все из них представляют собой малые лекарственные молекулы синтетической природы[5]. На фоне необъятного синтетически доступного химического пространства потенциальных лекарственных веществ ежегодное пополнение списка на 2–3 десятка новых лекарств представляет собой явление крайне редкой природы и как следствие влечет за собой высокую неопределенность, которая стоит перед медицинским химиком, в выборе стартовых точек и пути навигации в ходе дизайна потенциальных лекарственных веществ. В связи с этим наиболее общим решением проблемы размерности химического пространства в рамках задач по дизайну потенциальных лекарственных веществ является определение сценария разработки, позволяющее сузить область химического пространства, подлежащую рассмотрению.

1.1.2 Основные сценарии разработки потенциальных лекарственных веществ

Если в предыдущем столетии преобладали методы дизайна потенциальных лекарственных веществ, основанные на знании о структуре лигандов (LBDD, *ligand-based drug design*), то начиная с конца XX века наблюдается бурный рост доли программ по разработке лекарств, дизайн которых основывается на знании о структуре мишени (SBDD, *structure-based drug design*). Экспоненциальный рост числа записей со структурной информацией о строении мишеней и их комплексов с низкомолекулярными веществами (со-кристаллами) на сайте Банка структурных данных белков (PDB, *Protein Data Bank*) [7] в последние десятилетия

[5] Помимо малых лекарственных молекул синтетической природы в последние годы список активно пополняется моноклональными антителами и иными объектами преимущественно биологической природы.

позволяет использовать SBDD-парадигму для большинства современных программ по разработке потенциальных лекарственных веществ. Влияние LBDD подходов ещё сохраняется в определенных семействах молекулярных мишеней, таких как мембранные рецепторы, ввиду сложности получения структурной информации об их строении, однако же и оно постепенно снижается на фоне улучшения и развития как методов структурного анализа биополимеров, так и методов прогнозирования трёхмерной структуры биополимеров. В последние годы оптимизм в этой области был связан с появлением алгоритма AlphaFold2 [8], который позволяет прогнозировать трёхмерные структуры преимущественно белков с небывалой точностью. В то же время, среди профессиональных медицинских химиков сохраняется определенный скепсис относительно возможности использования моделей, полученных при помощи алгоритма AlphaFold2, в целях дизайна [9].

В ходе развития современной медицинской химии консолидировалось несколько базовых стратегий (методов) дизайна потенциальных лекарственных веществ на основе структуры мишени (включая, знания о структуре лиганд-белкового комплекса, иначе холо-структуре мишени или только структуре мишени без со-кристаллизованного лиганда — апо-структуре), которые можно классифицировать, как предлагается в таблице 1.

Таблица 1. Базовые стратегии дизайна потенциальных лекарственных веществ на основе структуры мишени

№	Стратегия	Входные данные	Инструментарий
1	*De novo* [10]	Апо-структура мишени	1. Молекулярный докинг, 2. Фармакофорный поиск на основе структуры мишени,
2	*De novo* [10]	Холо-структура мишени	1. Молекулярный докинг, 2. Фармакофорный поиск на основе структуры мишени, или на основе структуры лиганда, 3. Поиск по подобию формы лиганда
3	Hit-expansion [11]	Холо-структура мишени	1. Молекулярный докинг с удержанием структурного фрагмента, 2. Фармакофорный поиск на основе структуры лиганда, 3. Поиск по подобию формы лиганда
4	FBDD [12]		
5	Scaffold-hopping [13]		
6	Дизайн R-групп [14]		
7	Дизайн линкера		

Приведенные методы дизайна в первую очередь различаются по объему химического пространства, потенциального подлежащего рассмотрению, и перечислены в порядке уменьшения этого объема. Так, так называемые, *de novo* методы дизайна могут вообще не требовать каких-либо наперед заданных структурных элементов, ограничивающих химическое пространство [10]. В отдельных случаях возможно рассматривать в рамках *de novo* подхода использование диапазонов молекулярных дескрипторов или привилегированных фрагментов, характерных для класса молекулярных мишеней, если этот класс хорошо описан с точки зрения наличия активных малых молекул. Например, для ингибиторов протеиновых киназ характерно наличие ароматических гетероциклов, содержащих акцептор и донор водородной связи для эффективного связывания в hinge-регионе сайта связывания киназы. Поэтому медицинский химик будет руководствоваться этим знанием и использовать соответствующие гетероциклы в дизайне киназных ингибиторов, если его молекулярная мишень — протеиновая киназа. Эти же знания можно использовать для создания фармакофорной модели, в которую будут входить фармакофорные точки, характерные для киназных ингибиторов. Тем не менее, даже при использования подобных эвристик рассматриваемое химическое пространство в *de novo* стратегиях дизайна является всё ещё очень большим, несмотря на потенциальное преимущество в новизне создаваемых молекулярных структур.

Напротив, информация о конкретном соединении-хите, может представляться крайне полезной для сужения химического пространства. Если задаться задачей изучить химическое пространство вокруг соединения хита и отталкиваться от гипотезы, что некоторая часть структуры соединения хита отвечает за больший вклад в активность молекулы, то применяют стратегию *Hit expansion* (англ., расширение пространства соединений-хитов) [11], при этом размер фиксированной части структуры будет обратно пропорционален объему химического пространства, которое может подлежать рассмотрению в рамках кампании по дизайну молекулярных структур. Чаще всего такие гипотезы строятся на основе результатов экспериментов по молекулярному моделированию и степень достоверности этих экспериментов может значительно разниться.

Если же первичными соединениями-хитами являются малые по размеру молекулы (MW < 300 Да), или иначе фрагменты, а информация о связывании подтверждается методами структурного анализа биополимеров, такими как рентгеноструктурный анализ, криоэлектронная микроскопия или же методы фрагментного скрининга при помощи ЯМР-спектроскопии, то применяется стратегия дизайна потенциальных лекарственных веществ на основе знаний о связывании фрагмента с молекулярной мишенью (FBDD, *fragment-based drug*

design) [12]. Наличие такой информации позволяет не только ограничить пространство с точки зрения моделирования структур, которые должны будут содержать определенный фрагмент, но и значительно ограничить конформационное пространство и задать ограничения для молекулярного докинга, который должен будет обуславливаться на координаты атомов низкомолекулярного фрагмента, при этом эти координаты будут в значительно большей степени достоверны, чем те, что могут быть получены в результате молекулярного моделирования, выгодно отличая FBDD-стратегию от стратегии *Hit expansion* до тех пор, пока для последней соединение-хит не будет охарактеризовано вышеперечисленными методами структурного анализа биологических макромолекул.

Более узкий класс задач по молекулярной диверсификации и расширению химического пространства вокруг соединения-хита относится к трем выделяемым отдельно сценариям: *scaffold-hopping*, дизайн R-групп и дизайн линкера. Несмотря на отсутствие русскоязычного перевода для понятия *scaffold-hopping,* на практике смысл этого подхода отражается в замене одного скаффолда в молекулярной структуре на другой [13]. Такая техника позволяет разнообразить химическое пространство вокруг соединения-хита, а также радикально изменить физико-химические свойства ряда соединений. Что касается дизайна R-групп (заместителей), то такая менее инвазивная техника в основном призвана улучшить фармакокинетический профиль хемотипа и его активность, при сохранении большей части структуры неизменной [14]. Самым известным способом исполнения сценария, соответствующего дизайну R-групп, является техника, носящая название *halogen scan*, смысл которой заключается в последовательном переборе одного или нескольких галогеновых заместителей (чаще всего -F и -Cl заместители) замещенных ароматических циклов в рамках CCA- или CCC-оптимизации (CCA, соотношение структура-активность; CCC, соотношение структура-свойство). Последний выделенный сценарий, представляющий собой дизайн линкера, является несколько более разнообразным, исходя из того, для чего стоит искать линкер. Если линкер предстоит моделировать для несвязанных фрагментов, каждый из которых был получен в результате описанных выше экспериментов по структурному анализу, то такой сценарий стоит относить к продвинутым вариантам FBDD-стратегии моделирования [15]. Если два фрагмента являются более крупными, или даже полноценными лигандами, один из которых представляет лиганд E3-убиквитин лигазы, а второй — лиганд целевой молекулярной мишени, то дизайн линкера становится основной задачей структурного моделирования химерных (комбинированных) соединений, индуцирующих протеолиз мишени (PROTAC, *proteolysis targeting chimera*) [16]. Если же объединяемые линкером подструктуры относятся к одной и той же молекуле, то результатом становится циклизация

молекулы. В таком случае, чаще всего прибегают к макроциклизации молекулярных структур, что в последние годы приобрело особую популярность в дизайне ингибиторов протеиновых киназ и вирусных протеаз, благодаря лучшим фармакокинетическим свойствам макроциклов и более выгодному позиционированию макроциклических структур с точки зрения патентованного химического пространства [17].

Таким образом мы обозначили базовые сценарии моделирования структур потенциальных лекарственных веществ, которые рассматриваются в рамках компьютеризированного дизайна потенциальных лекарственных веществ (англ., *computer-assisted drug design*, CADD), и на которые имеет смысл опираться и ссылаться при создании любого ПО, связанного с моделированием структур потенциальных лекарственных веществ.

1.2 Концепция генеративной химии и использование искусственного интеллекта в дизайне потенциальных лекарственных веществ

1.2.1 Понятие о генеративной химии

Несмотря на то, что разработка лекарственных препаратов — это чрезвычайно сложный процесс, требующий тщательного учета множества важных критериев и серьезных интеллектуальных и экономических вложений, идея минимального вмешательства специалиста в процесс моделировании структур потенциальных лекарственных веществ и постановка этого процесса в потоковый режим давно интересовала профессиональное сообщество медицинских химиков. Потребности растущих крупной фармацевтической и биотехнологической промышленностей в реализации этого интереса на практике в купе с ростом биотехнологического сектора способствовали тому, что с начала 90-х годов сложились основания для зарождения новой области химического знания, которое впоследствии начало называться *генеративной химией*.

Существует два взгляда на определение понятия “генеративная химия”. Эти два взгляда сходятся на том, какая задача должна решаться этой областью научного и прикладного знания, а именно **задача автоматизированной генерации молекулярных структур веществ под заданный профиль свойств**. Этими веществами могут быть потенциальные лекарства, материалы, агрохимикаты и иные важные для современной экономической деятельности группы химических веществ. И в то же время те же два взгляда расходятся во мнении, каким инструментарием эта задача должна решаться. Первый взгляд состоит в том, что генеративная химия — это область современного прикладного искусственного интеллекта (ИИ) и её задачи решаются при помощи современных методов ИИ, в первую очередь методами, основанными

на, так называемом, глубоком обучении искусственных нейронных сетей [18]. Этот взгляд, как минимум, историчен, поскольку сам термин “генеративная химия” (*generative chemistry*, англ.) — относительно нов и не появляется в профильной литературе до 2020 года (см. рис.1). Впервые термин был введен в статье А. Жаворонкова и его коллег [19] и подробно разобран в другой статье [20] почти того же коллектива авторов в том же 2020 году. В обеих статьях авторы придерживаются упомянутого взгляда на инструментарий генеративной химии. Второй же взгляд на инструментарий, которым может решаться задача генеративной химии, заключается в том, что он может трактоваться более широким способом, а именно распространяться на любые методы и алгоритмы, в том числе и самые примитивные, в основе которых лежит идея об автоматической генерации химических структур веществ с заданными свойствами, и не ограничивается исключительно методами, основанными на “глубоком обучении”. Автор настоящей квалификационной работы придерживается последнего взгляда с более широкой трактовкой понятия “генеративной химии”, поскольку считает, что ограничение лишь современными инструментами генерации отрезает широкий пласт ранних генеративных моделей, которые, тем не менее, решают задачу генеративной химии, пусть и на более примитивном и ограниченном уровне, не прибегая к методам искусственного интеллекта[6]. В таком случае, понятие генеративной химии будет синонимично другому понятию — “автоматизированный химический дизайн”, который был детально рассмотрен и классифицирован по уровням автоматизации генерации идей и автоматизации принятия решений в статье Б. Голдмана и коллег [22].

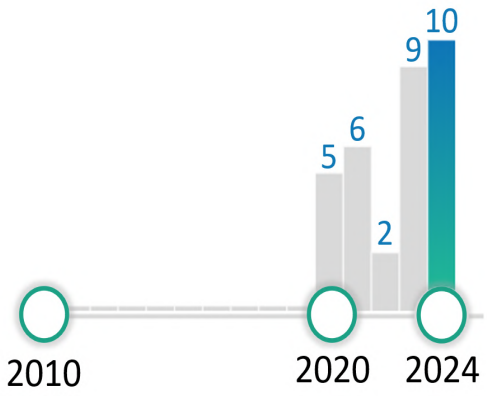


Рисунок 1. Количество статей в рецензируемых журналах, индексируемых на ресурсе PubMed по запросу “generative chemistry” в названиях, абстрактах и полных текстах статей за период с 2010 по 2024 гг. Первое упоминание термина “generative chemistry” датируется 2020 годом.

[6] Данный взгляд нашел свое отражение в обзорной статье коллектива авторов из группы компаний Insilico Medicine, включая автора настоящей квалификационной работы [21].

1.2.2 Ранние генеративные модели

Уже в 90-ые годы XX века начали появляться первые алгоритмы автоматической генерации и многопараметрической оптимизации структур потенциальных лекарственных веществ, которые представляли собой генераторы виртуальных синтетических комбинаторных библиотек, модификаторы R-групп и генетические алгоритмы [23]. Отдельные попытки были сделаны и в создании итеративного, поатомного дизайна в трехмерных координатах [24].

Особняком стоят две простейшие генеративные модели RECAP (Retrosynthetic Combinatorial Analysis Procedure) [25], созданная в конце 90-х годов, и BRICS (Breaking of Retrosynthetically Interesting Chemical Substructures) [26], предложенная на 10 лет позже, но наследующая идеологию RECAP. Эти модели сфокусированы на генерации молекулярных структур веществ с более высокой ожидаемой синтетической доступностью. Основной идеей этих двух моделей является генерация новых молекулярных структур путем комбинаторной рекомбинации фрагментов синтезированных молекулярных структур, причем эти фрагменты получаются путем квази-ретросинтетических реакций, записанных на языке подструктур SMARTS. На рис. 2. А. изображены генетические связи 16 типов фрагментов, генерируемых при помощи алгоритма BRICS, и на рис. 2. Б. — пример разбиения алгоритмом молекулярной структуры киназного ингибитора сорафениба (sorafenib); так, например, разрыв ретросинтетически значимой амидной связи, приводит к фрагментам L1 (ацил) и L2 (амин), разрыв связи, соответствующей простым эфирам фенолов — к фрагментам L14 (арил) и L3 (алкилированный азот), а фрагмент L7 (алкенил), замыкающийся сам на себя, соответствует трансформу реакции Виттига.

Сама же комбинаторная рекомбинация фрагментов производится тем же набором реакций, но уже в прямом направлении. Тем самым авторы метода добиваются более высокой ожидаемой синтетической доступности полученных в результате такой рекомбинации молекулярных структур. Алгоритм BRICS доступен для использования из хемоинформатической библиотека с открытым исходным кодом RDKit [27]. За фрагментацию в рамках алгоритма ответственен метод BRICSDecompose.

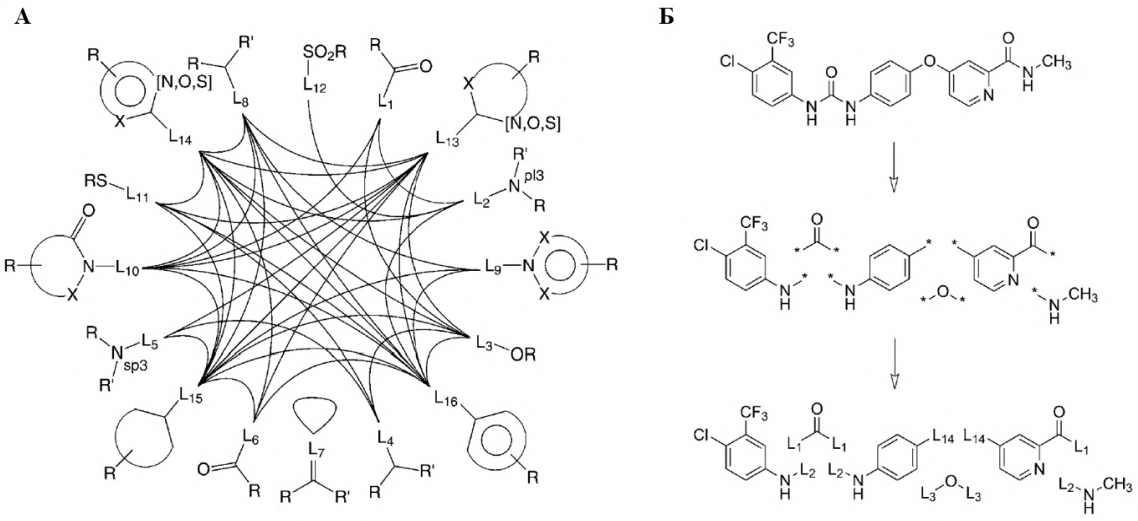


Рисунок 2. А. Реакционные типы фрагментов, получаемых при помощи алгоритма BRICS. **Б.** Схематичное описание принципа квази-ретросинтетической фрагментации при помощи алгоритма BRICS.

1.2.3 Современные генеративные модели, основанные на алгоритмах глубокого обучения

Расцвет технологий искусственного интеллекта после продолжительного периода стагнации, иначе именуемого как “зима искусственного интеллекта”, был вызван появлением так называемого “глубокого обучения” искусственных нейронных сетей (далее ГО) в первой половине второго десятилетия XXI века [28]. На волне оптимизма, в первую очередь связанного с радикальным улучшением качества генерации изображений, новые архитектуры ИИ на основе ГО стали применяться и для решения задач в области “наук о жизни” [29], в том числе и для дизайна потенциальных лекарственных веществ. Параллельно с развитием алгоритмов ГО развивалась и хемоинформатика — область применения методов информатики для решения химических проблем[7]. В частности, широкое развитие и применение получила ранее уже упомянутая хемоинформатическая библиотека с открытым исходным кодом RDKit на основе языка программирования Python [27].

Впервые генерация молекулярных структур при помощи ГО была представлена в препринте Р. Гомеза-Бомбарелли и коллег в 2016 году [31][8]. Архитектурное решение, использованное для генерации, представляло собой вариационный автокодировщик (VAE, *variational autoencoder*) — одну из разновидностей глубоких нейронных сетей. Позднее, в

[7] По определению И. Гастайгера [30]

[8] Препринт от 2016 года на ресурсе arxiv.org был переиздан в статье *ACS Central Science*

2017–2018 гг., специалистами из группы компании Insilico Medicine были предложены несколько архитектур на основе ГО, позволяющих генерировать новые молекулярные структуры соединений-хитов [32–34]. Так на примере, Jak3 тирозиновой киназы, была продемонстрирована принципиальная возможность сгенерировать молекулярную структуру ингибитора киназы при помощи улучшенной архитектуры автокодировщика (*Entangled Conditional Adversarial Autoencoder*), несмотря на скромное значение IC_{50} (6.73 μM) [32].

Однако, первый наглядный пример того, что ГО способно радикально ускорять процесс поиска потенциальных лекарственных веществ был продемонстрирован исследователями из группы компаний Insilico Medicine, в числе которых был автор настоящей диссертационной работы. Статья, опубликованная в 2019 году в журнале Nature Biotechnology, была посвящена дизайну низкомолекулярных ингибиторов DDR1 киназы — важной молекулярной мишени в терапии фиброза лёгких [35]. В ходе исследования (см. рис. 3) специалистам удалось в короткий срок (7 дней) подготовить обучающий набор данных, содержащий информацию об известных ингибиторах DDR1 киназы. Именно этот набор данных был использован ключевым генеративным алгоритмом, называемым моделью GENTRL (***GEN****erative* ***T****ensorial* ***R****einforcement* ***L****earning*), для обучения. Другие наборы данных были собраны для обучения самоорганизующихся карт (СОК) Кохонена [36], выполнявших функцию оценки генерируемых молекулярных структур. Помимо СОК награждение структур так же производилось на основе оценки их фармакофорного подобия известным ингибиторам DDR1 киназы. После почти двухнедельного генерационного цикла (12 дней) полученные ~30 000 молекулярных структур были подвергнуты различным процедурам пост-фильтрации, включая медхимические подструктурные фильтры (MCF, *medicinal chemistry filters*), фильтрацию по физико-химическим дескрипторам, ограничение размера кластеров, после процедуры кластеризации и др. После данной процедуры, которая заняла 2 дня, удалось отобрать 40 молекулярных структур, которые затем были отправлены в контрактную исследовательскую организацию (CRO, *contract-research organization*) WuXi Apptec. В свою очередь химики-синтетики из WuXi Apptec выполнили анализ синтетической доступности предложенных молекулярных структур и приоритезировали наиболее перспективные 6 структур, которые, в конечном итоге, и были синтезированы, и 2 из них продемонстрировали высокую активность в наномолярном диапазоне, что на 2-3 порядка лучше, чем активность ранее упомянутого ингибитора Jak3 киназы [32].

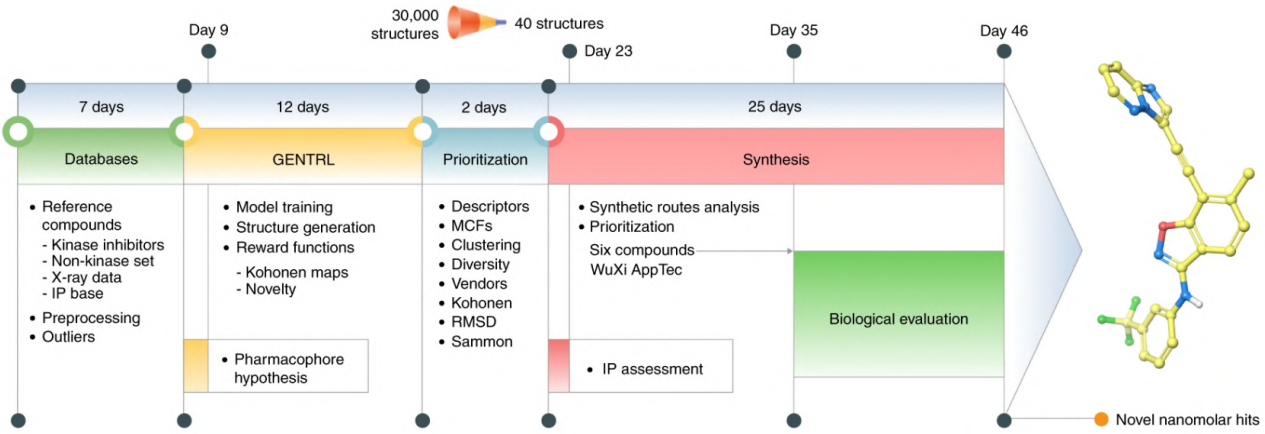


Рисунок 3. Схематичное описание процесса автоматизированного дизайна новых ингибиторов DDR1 киназы при помощи генеративной модели GENTRL [35].

В последнее время наблюдается экспоненциальный рост числа публикаций, эксплуатирующих идею использования различных архитектур ГО для решения тех (*de novo* генерация) или иных (оптимизация хемотипа) задач молекулярного моделирования потенциальных лекарственных веществ. Однако, настоящий обзор литературы не имеет целью обозреть весь тот огромный пласт публикаций, посвященных новым архитектурам генеративных моделей на основе ГО для поиска новых лекарств. Тем не менее, не смотря на то, что в профессиональном сообществе медицинских химиков и CADD-специалистов нередко складывается скептическое отношение к качеству подобных публикаций, учитывая что большинство из них не содержат информации об экспериментальной валидации результатов, стоит обратить внимание на относительно свежий обзор, опубликованный в июне 2023 г. нашей научной группой, цель которого — познакомить профессиональное сообщество коллег из области прикладного ИИ и МО с корпусом критических взглядов медицинских химиков на низкокачественные статьи в смежной области [21]. В качестве дидактического материала имеет смысл особое внимание обратить на выжимку критического характера (*Cliff notes*), приведенную в качестве заключения к упомянутому обзору, и рекомендации будущим авторам новых генеративных моделей.

1.2.4 Платформенные решения для выполнения задачи генеративной химии

Появление интегрированных, удобных в пользовании и не требующих навыков программирования платформенных решений для моделирования структур новых потенциальных лекарственных веществ было вопросом времени. Более того, в рамках подобных платформенных решений легче применять модульное строение и снабжать платформу новыми модулями, обеспечивающими новое функциональное наполнение.

Платформа Chemistry42 [37], ставшая доступной широкому кругу пользователей в августе 2020 года, была одним из первых подобных решений на формирующемся рынке программного обеспечения для решения задач генеративной химии. Поскольку создание подобной платформы является крайне трудоемким процессом и подразумевает вовлечение не одного десятка специалистов, включая медицинских химиков, хемоинформатиков, программистов широкого спектра и проектных менеджеров, то в настоящей диссертационной работе будут рассмотрены только те аспекты, за которые в ходе работы над платформой непосредственно отвечал автор настоящей квалификационной работы. Подробно функционал платформы будет рассмотрен в разделе 3.1, в то время как избранные главы, касающиеся модулей оценки синтезируемости молекулярных структур и оценки соответствия структурным трендам медицинской химии будут рассмотрены в разделах 2.2, 3.2 и 2.1 и 3.3 соответственно.

В 2021 году компания Iktos, специализирующаяся в области генеративного ИИ, анонсировала два своих основных продукта Makya [38] и Spaya [39]. Первый продукт представляет собой платформу генеративной химии, в то время как второй — платформу по выполнению автоматизированного ретросинтетического анализа. К сожалению, в свободном доступе пока не было представлено полноценного описания обеих платформ. В 2022 году корпорация Schrodinger, лидер в области создания инструментов CADD, предложил собственное решение для выполнения автоматизированной оптимизации молекулярных структур под названием AutoDesigner [40]. Основным преимуществом AutoDesigner является интеграция с продвинутым движком по оценке свободной энергии связывании FEP+ (*free energy perturbation*, англ.), благодаря чему удается с высокой точностью прогнозировать активность генерируемых молекулярных структур и отбирать на синтез наиболее перспективные кандидаты [41]. Тем не менее, платформа AutoDesigner не призвана решать задачи, связанные с *de novo* генерацией молекулярных структур.

В 2023 году компания Sumulation Plus, специализирующаяся в первую очередь на решениях для прогнозирования ADMET-свойств потенциальных лекарственных веществ, выпустила на рынок собственную платформу генеративной химии (*The AI-driven Drug Design (AIDD) platform*, англ.) [42]. Основной фокус платформы направлен на итеративную мультипараметрическую оптимизацию с учетом широкого спектра прогнозируемых фармакокинетических и фармакодинамических характеристик и синтетической доступности. В качестве модели синтетической доступности авторы платформы вдохновлялись широко известным SA Score [43], разработанным сотрудниками компании Novartis. Модель

синтетической доступности, лежащая в основе SA Score, будет подробно разобрана в разделе 1.4.3.3.4 настоящего обзора литературы.

Стоит отметить, что если ранее создание платформ генеративной химии было делом малых и средних фармацевтических и биотехнологических компаний, то в последнее время фиксируется тренд на создание собственных платформенных AIDD-решений у ведущих мировых фармацевтических компаний. Так, например, в конце 2023 года компания Merck анонсировала свою технологическое решение, а в 2024 году сразу две компании, Novartis и AstraZeneca, предложили свои платформы, которые они развивают внутри компании не для широкого пользования. И если AIDISSON [44] (Merck) по описанию похож на существовавшие ранее платформы генеративной химии, то MicroCycle [45] (Novartis) и PIP [46] (AstraZeneca) в большей степени сфокусированы на том, чтобы максимально оптимизировать DSTA-цикл (***D**esign, **S**ynthesis, **T**esting, **A**nalysis*, англ.) разработки потенциальных лекарственных веществ путем интегрирования непрерывно-получаемых данных о синтезе, биологических тестированиях в процесс моделирования новых молекулярных структур при помощи технологий искусственного интеллекта.

Бурный рост разработки новых генеративных моделей и платформенных решений позволяет говорить о том, что постепенно на замену CADD-методам приходит AIDD парадигма моделирования структур потенциальных лекарственных веществ (AIDD, *Artificial intelligence-assisted drug design* — дизайн лекарственных молекул при помощи искусственного интеллекта). Тем не менее, в профессиональном сообществе, в особенности в последнее время, активно обсуждается вопрос об эффективности такого парадигмального сдвига [47]. Поскольку многие платформы генеративной химии существуют уже не один год, то у многих возникает мнение, что пора начинать критически относиться к активности в этой созревающей отрасли, хотя еще два-три года назад все предоставляли ей *carte blanche* [48]. Объективным мерилом эффективности платформенных решений в области генеративной химии являются результаты доклинических и клинических испытаний потенциальных лекарственных веществ, вышедших из платформы, то есть количество доклинических кандидатов (*preclinical candidate*, PCC) и клинически исследуемых веществ, находящихся в различных фазах. Стоит сразу отметить, что пока ни одно потенциальное лекарственное вещество, будучи родом из AIDD-платформ, не прошло полный цикл клинических испытаний и не было зарегистрировано национальным регулятором. Наивысшее достижение, которое было на данный момент продемонстрировано платформой генеративной химии — это успешное завершение IIa фазы клинических испытаний ингибитором TNIK киназы (регистрационный номер КИ: NCT05938920) [49,50]. Сам ингибитор был получен благодаря упомянутой ранее и

рассматриваемой в рамках настоящей диссертационной работы платформе генеративной химии Chemistry42 [51].

Тем не менее, даже несмотря на наличие конечных объективных показателей эффективности платформ генеративной химии, никто не отрицает факт того, что есть и объективные промежуточные показатели предельной важности. Например, это доля молекулярных структур от общего набора молекулярных структур сгенерированных алгоритмами ИИ, отобранных на синтез и успешно синтезированных в лабораторных условиях (*synthesis success rate*) [22], что является своеобразным аналогом другой важной промежуточной метрики *hit rate*, которая отражает долю молекул, для которых предсказание активности совпадает с экспериментальным фактом. Очевидно, что химическое пространство, созданное при помощи ИИ, которое сходно по наполнению с тем, что было синтезировано ранее, окажется более синтетически доступным, нежели то, что не похоже на прецедентные случаи синтеза. С другой же стороны, фармацевтическая индустрия заинтересована в новых хемотипах, выходящих за рамки запатентованного химического пространства. В связи с этим стоит рассмотреть главную дилемму, стоящую перед алгоритмами генеративной химии, заключающуюся в существовании двух противоположно направленных требования к результатам их работы:

1. Генерируемые молекулярные структуры должны обладать структурной новизной и выходить за рамки хорошо исследованного и запатентованного химического пространства;
2. Генерируемые молекулярные структуры должны обладать высокой синтетической доступностью.

Оба этих требования действуют, как правило, в противоположном направлении в процессе дизайна. Очевидно, что если молекулярная структура обладает высокой синтетической доступностью, что может означать, что она может быть получена за минимальное число стадий из коммерчески доступных и недорогих стартовых соединений, используя надежные синтетические методы, то высок риск того, что эта молекулярная структура была уже ранее кем-то отправлена в синтетическую лабораторию и воплощена в веществе. И наоборот, чем более новой со структурной точки зрения является молекулярная структура, тем ниже вероятность того, то для нее может быть найден лаконичный путь синтеза с высоким шансом на успех. Таким образом, диалектическая картина генеративной химии требует уделить особое внимание пониманию того, как и куда развивается с точки зрения структурного усложнения медицинская химия и как сбалансировать это усложнение надежными методами

моделирования синтетической доступности. Нахождение этого баланса, помимо других факторов, является важнейшим условием для успешного функционирования любой платформы генеративной химии.

1.3 Структурные тренды и эволюция медицинской химии

Более полутора веков назад Чарльз Дарвин и Альфред Уоллес сформулировали основные принципы революционной на тот момент теории эволюции как самоорганизующегося, в целом энтропийно-управляемого, но не хаотического процесса естественного отбора и адаптации живых организмов к окружающей среде. С тех пор был достигнут огромный прогресс практически во всех сферах жизни, науки и технологий. Заметный рост во многих биологических дисциплинах позволил ученым сформулировать однозначные механистические и философские объяснения «как и почему» произошла эволюция, и как она подпитывается рядом жизненно важных импульсов. В противоположность технологическому и промышленному прогрессу была представлена альтернативная гипотеза биологической природы, которая постулирует основные принципы сценария деволюции параллельно с эволюцией или вместо неё. Онтологически мы можем применить фундаментальные представления об эволюции и деволюции к медицинской химии и разработке лекарств через призму ключевых факторов, доминирующих в этих областях. Химический «большой взрыв» был ценным и универсальным источником практически неисчерпаемой вселенной, обильно населенной астрономическим количеством гипотетических структур, обладающих подобием лекарствам ($\sim 10^{60}$) [52]. Хотя было зарегистрировано более 135 миллионов молекул, корпоративные коллекции содержат всего ~8–10 миллионов уникальных молекул (исключая строительные блоки), и большинство из них представляют собой собрания «старых» кластеров, которые, как правило, не представляют интереса для современных проектов по дизайну потенциальных лекарственных веществ. Эти соединения часто выходят за рамки текущих критериев новизны и не соответствуют основным тенденциям, наблюдаемым в современной разработке лекарств. Бум высокопроизводительного скрининга (ВПС) существенно истощил доступные запасы химических библиотек, в результате чего огромное количество молекул остаются в хранилищах, с высокой вероятностью превращаясь в продукты деградации без должного и регулярного контроля качества. В этой нише подавляющее большинство скаффолдов и хемотипов в настоящее время считаются непривлекательными и имеют ограниченный потенциал, и, словно динозавры, эти скаффолды исчезают один за другим из практики.

Разнообразие и филогения биологических мишеней тесно связаны с медицинской химией. Эти дисциплины слились в единый симбиотический континуум и долгое время развивались вместе. Как правило, впечатляющие взрывы в медицинской химии отражают биологические прорывы и наоборот. То есть под новый класс биологических мишеней формируется целый тренд в медицинской и синтетической химии, как это, например, произошло с ингибиторами протеиновых киназ после осознания того факта, что протеиновые киназы как биологические мишени пригодны для разработки лекарств. И наоборот, если синтезируется элегантная и новая молекула, несущая гармоничный ансамбль потенциальных точек связывания, или появляется новый мощный метод синтетической химии, позволяющий синтезировать структурно новые молекулы, можно с большой вероятностью найти подходящую мишень (обратная задача дизайна потенциальных лекарственных веществ). Наиболее наглядно это происходило в связи с внедрением реакции кросс-сочетания в практику медицинской химии, что позволило создавать лекарства ранее просто недоступные по синтетическим соображениям, а теперь современную медицинскую химию нельзя помыслить без реакции Сузуки-Мияуры или реакции Соногаширы. Равно как и появление цикл-замыкающего метатезиса (RCM, *ring-closing metathesis*) способствовало появлению макроциклических киназных и протеазных ингибиторов. Эволюция биологических мишеней, под которой мы понимаем постепенный сдвиг рамки восприятия пригодности биологических мишеней от статуса “непригодный” к статусу “пригодный”, смещает фокус медицинской химии в первую очередь в сторону сложных структур, подобных природным продуктам, в основном из-за значительного прогресса, достигнутого в области взаимодействий между белками (ВМБ), участвующих в сигнальных путях заболеваний [53]. В настоящее время ученые предполагают, что они идентифицировали более 400 000 ВМБ, охватывающих разнообразные связывающие борозды с объемами 800–2 000 $Å^3$ и сайты связывания объемом 250–1 000 $Å^3$. Пожалуй, самым ярким примером последних лет, иллюстрирующим эту картину, является одобрение ленакапавира (lenacapavir, рис. 4) FDA в 2022 году для лечения ВИЧ-инфекции.

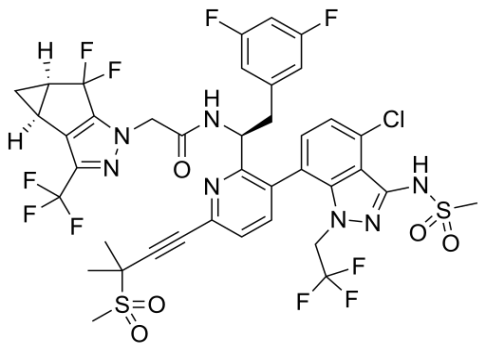


Рисунок 4. Структурная формула ленакапавира, одобренного для лечения ВИЧ инфекции в 2022 году FDA. Структура наглядно иллюстрирует выход за пределы химического пространства, подчиняющихся правилам Липински [54].

Неудивительно, что компании развернули масштабные программы по созданию библиотек, ориентированных на ВМБ, содержащих большое количество новых алифатических и гетероалифатических систем, спиро- и хиральных центров, мостиковых и конденсированных колец для достижения лучшего охвата 3D-химического пространства. Однако ежегодно на рынок выводится лишь около 15 новых молекулярных сущностей, что подчеркивает стагнацию фармацевтической индустрии, главным образом из-за высокого уровня отсева проектов и растущих барьеров в области интеллектуальной собственности. Таким образом, перед химиками, работающими в области медицинской химии, нависает огромная ответственность, так как они вовлечены в этот нетривиальный процесс. В настоящее время требуется значительное время для разработки новых молекул, соответствующих преобладающим критериям, которые постепенно отходят от оптимальных значений, установленных традиционными правилами Липински [54] и связанными индексами. Поэтому мы рассматриваем это как проявление «химической сингулярности». Мы проводим аналогию (аллегорию) между гравитационной сингулярностью черной дыры и ВПС. В начале века технология ВПС развивалась, становилась более мощной и коммерчески доступной для многих исследовательских организаций. В тот период было доступно огромное количество молекул в коллекциях поставщиков, как материя, конденсированная вокруг черной дыры. В течение следующих двух десятилетий многие из этих молекул были проанализированы в ходе различных кампаний ВПС. Кроме того, химики и специальные фильтры переместили значительное количество соединений за пределы пространства, пригодного для создания лекарств априори. В результате количество этой реальной «химической материи» значительно уменьшилось и коллапсирует к “химической сингулярности” как одномерной точке, содержащей огромную химическую массу. *De facto*, на данный момент эти молекулы находятся за пределами “горизонта событий”, и относительно небольшое количество

соединений может его покинуть. Новые соединения, очевидно, имеют более сложную структуру и разительно отличаются от старых химических сущностей (см рис. 5).

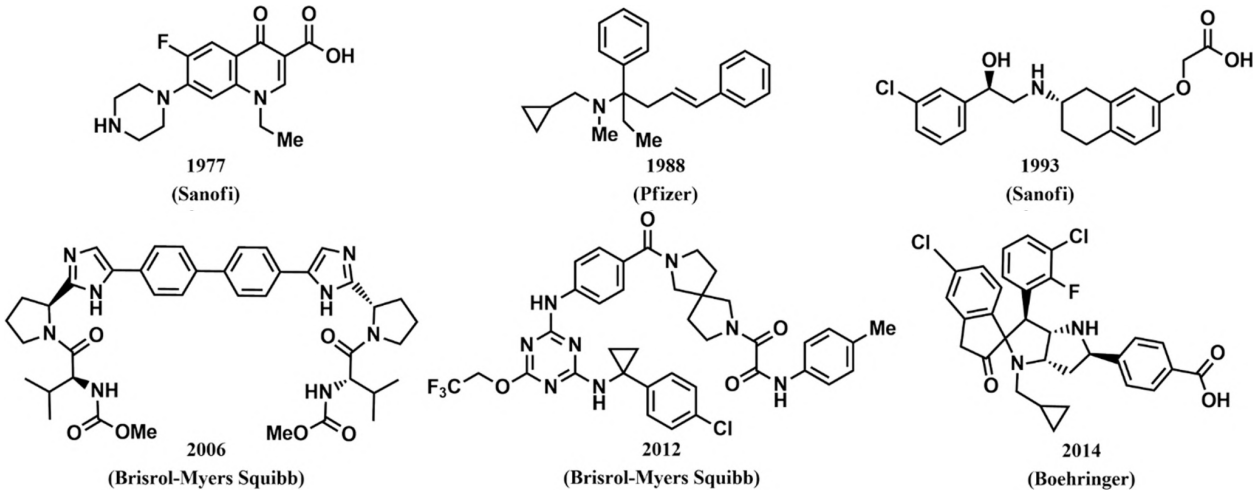


Рисунок 5. Структурные формулы веществ, запатентованных крупнейшими фармацевтическими компаниями. В верхнем ряду приведены примеры веществ, разработанных в прошлом столетии, в нижнем ряду — примеры веществ, разработанных в текущем столетии.

Многие современные программы разработки лекарств начинаются с *de novo* дизайна, и неудивительно, что “химическая материя” не может быть быстро пополнена, поскольку скорость синтеза значительно уступает скорости ВПС. Учитывая это, эволюция медицинской химии становится более очевидной.

Попытки определить структурные тренды современной медицинской химии наблюдались с конца прошлого столетия. В этом смысле, упомянутая работа К. Липински [54] стала первым пробным камнем в определении того, что такое “перспективное” химическое пространство. Эта же работа, по иронии судьбы, как это бывает с первыми в своем роде научными концепциями[9], стала своего рода тормозом как в развитии *понимания химического пространства молекул,* обладающих подобием лекарствам, так и в понимании *развития такого пространства*. В результате, начиная с середины 2010-х годов, то есть спустя почти 15–20 лет после публикации работы Липински, начинают выходить первые аналитические работы, о покидании новыми drug-like молекулами области, очерченной “правилами пяти” (*“beyond rule of five”*). В ряде таких работ стоит выделить статьи от исследователей из компаний AstraZeneca [55], Novartis [56,57], Bayer [58] и AbbVie [59]. Авторы этих работ в

[9] Здесь уместно сравнение правил Липински, например, с птолемеевской концепцией геоцентрической системы мира. Последняя затормозила развитие астрономии на многие века.

целом сходятся на том, что использование классических подходов к определению подобия лекарствам (*drug-likeness*) существенно и во многом необоснованно ограничивает арсенал современной медицинской химии: и актуальные значения молекулярного веса (MW, *molecular weight*, англ.) или показателя липофильности (logP) значимо превосходят пороги из “правил пяти”, тогда как развитие методологий дизайна лекарственных молекул на основе направленного протеолиза (PROTAC) вообще приводит к полному отказу от классических представлений о подобии лекарствам [60].

Особняком стоит статья Ф. Ловеринга [61] из компании Wyeth (ныне часть Pfizer), где автор, не в противовес правилам Липински, а скорее в дополнение, предложил новый на тот момент структурный дескриптор Fsp^3, — долю (фракцию) атомов углерода, находящихся в sp^3-гибридизации, — связав, на основании статистического анализа, более высокие значения Fsp^3 с повышением вероятности малой молекулы пройти клинические испытания. Несмотря на новый взгляд на структурные тренды, которые можно наблюдать при помощи дескриптора Fsp^3, невозможность различия sp^3-гибридизированных атомов углерода, находящихся в длинных углеродных цепочках и тех, что находятся в карбо- и гетероциклах, представляется существенным недостатком дескриптора. Это недостаток мы попытались обойти при разработке нового дескриптора MCE-18, описывающего структурные тренды современной медицинской химии, что непосредственно вошло в исследовательский корпус настоящей диссертации (см. разд. 3.3).

1.4 Моделирование синтетической доступности структур потенциальных лекарственных веществ

1.4.1 Предварительные замечания о понятии синтетической доступности

Несмотря на то, что термин *“синтетическая доступность”* довольно широко используется в теории и практике компьютеризированного дизайна лекарственных молекул, стоит в первую обозначить два важных замечания относительно понятия о *“синтетической доступности”*:

1. В англоязычной литературе, как основном источнике знаний о современной медицинской химии, отсутствует устоявшееся и общепринятое определение того, что такое *“синтетическая доступность”* [62].

2. В связи с этим наличествует путаница между несколькими семантически похожими терминами, которые можно найти в профильной литературе, которые на русский язык переводятся часто тоже похожим неорганизованный образом. К числу этих терминов можно отнести следующие:

synthetic accessibility — то, что, на наш взгляд, и стоит переводить как *“синтетическая доступность”*;

synthetic feasibility — то, что часто переводят как *“синтетическая доступность”*, но, по нашему мнению, стоит переводить скорее как *“синтетическая осуществимость”* или *“синтетическая реализуемость”,* или *“синтезируемость”*;

synthesizability — можно перевести как *“синтезируемость”,* и в таком случае он будет полным синонимом вышеупомянутой *“синтетической осуществимости”* ;

synthetic complexity — *“синтетическая сложность”* или *“сложность синтеза”*, абстрактное понятие [63], косвенно связанное с понятием синтетической доступности, но, тем не менее, имеющее отношение скорее к трудоемкости синтетической кампании, которая, например, может быть приближена числом стадий синтеза, чем к оценке осуществимости синтеза некоторой молекулы;

molecular complexity и *structural complexity* — молекулярная или структурная сложность [64,65] — непосредственно исчисляемая в молекулярных и структурных дескрипторах характеристика молекулярной структуры, основанная на предположении, что некоторые структурные элементы или характеристики (эвристики) молекулы могут быть ассоциированы со сложностью в осуществлении синтеза данной молекулярной структуры. К таким структурным элементам относят асимметрические атомы углерода, спиро-сочленения, конденсированные циклы и др., а к таким характеристикам молекулы относят разветвленность и большой молекулярный вес.

3. Нередко авторы научных работ, описывающих новые методы моделирования синтетической доступности, преподносят новые термины, касающиеся тематики, и не пытаются осмыслить их относительно уже существующих. Так, например, произошло с термином *synthetic complexity* [63], который не был в должной мере отрефлексирован относительно более распространенного термина *synthetic accessibility.*

В связи с этим, мы считаем необходимым в первую очередь определить понятие *“синтетической доступности”*. Более того, в настоящем исследовании мы, в том числе задаемся целью впервые в профильной литературе представить математическое описание понятия о “синтетической доступности” на языке теории вероятности и позиционировать это понятие относительно других синонимичных понятий.

1.4.2 Вероятностное определение синтетической доступности и следствия из него

Если задаться целью разграничить понятия синтетическая осуществимость (*synthetic feasibility*) и синтетическая доступность (*synthetic accessibility*), то представляется, что под синтетической осуществимостью стоит понимать **принципиальную** возможность синтезировать или не синтезировать вещество в рамках текущего технологического развития, используя весь корпус знаний об органическом синтезе. При этом неважно какие ресурсы могут быть направлены на проект по синтезу. Например, полные синтезы структурно сложных веществ биологического происхождения может выполняться целой группой химиков-синтетиков на протяжении нескольких лет. То есть, формулируя вопрос о синтезируемости (синтетической осуществимости) некоторого молекулярного объекта, мы на самом деле получаем закрытый вопрос, ответами на который могут быть только *“да, молекула синтезируема”* и *“нет, молекула не синтезируема”*. В то же время, кажется, что такая постановка вопроса ведет к тому, что окончательный ответ можно получить только на практике, перепробовав абсолютно все возможности по синтезу молекулы, что, на первый взгляд, выглядит бесполезно, но, с другой стороны, идеально подходит под описание бернуллиевской случайной величины.

Итак, синтетическая осуществимость (далее СО) может быть определена как свойство молекулы, которая может быть либо синтезируемой (исход 1), либо несинтезируемой (исход 0), или, другими словами, синтетически реализуемой (1) или нереализуемой (0), с учетом всего текущего синтетического знания и технологий, всех возможных синтетических планов, всех доступных исходных материалов, затрат, труда и времени, которые могут быть выделены на синтетическую кампанию для получения молекулы в виде синтезированного соединения. Таким образом, СО может рассматриваться как бернуллиевская случайная величина с двумя возможными исходами:

$$\mathrm{CO}(w) = \begin{cases} 0, & \text{если молекула синтезируема;} \\ 1, & \text{если молекула не синтезируема.} \end{cases}$$

В таком случае, синтетическую доступность (далее СД) можно определить как вероятность (оценку) p того, что бернуллиевская случайная величина СО примет значение 1. Тогда функция распределения случайной величины СО определяется следующим образом:

$$F_{\mathrm{CO}}(x) = \begin{cases} 0, & \text{при } x < 0; \\ 1 - \text{СД}, & \text{при } 0 \le x < 1; \\ 1, & \text{при } x \ge 1. \end{cases}$$

Дополнение к СД q представляет собой вероятность того, что соединение не будет

синтезируемым при текущих условиях (вероятность неудачи, СО (ω) = 0):

$$q = 1 - p = 1 - \text{СД}.$$

Это учитывает присущую неопределенность и ограничения в синтетической химии, отражая вероятность того, что, несмотря на теоретическую доступность, практический синтез может оказаться неосуществимым.

Поскольку СД определяется как бернуллиевская вероятность *p*, все свойства и следствия вероятности верны для СД, включая закон полной вероятности:

$$\text{СД} = \sum_{i=1}^{n} P(\text{СО} = 1 \mid C_i) \cdot P(C_i),$$

$$\text{СД}_i = P(\text{СО} = 1 \mid C_i),$$

что учитывает множество различных исчислимых (*n*) наборов условий (C_i), в которых можно вычислить связанные с условием значения СД_i. Такие условия включают различные доступные наборы исходных материалов, синтетические планы для оценки, доступные синтетические методы и технологии в лаборатории, трудовые ресурсы, бюджет и время, которые могут быть выделены на синтетическую кампанию и т. д.

Учитывая то, что, исходя из предложенного вероятностного описания значений синтетической доступности как оценки вероятности того, что молекула является синтетически осуществимой, широко применяемое сочетание слов “оценка синтетической доступности” мы считаем тавтологией, в то время как термины “оценка синтезируемости” и “синтетическая доступность” являются логичными и синонимичными. В отношении различных алгоритмов и подходов по оценке синтезируемости молекул мы будем использовать термин “метод моделирования синтетической доступности” (**ММСД**), поскольку каждый из них имеет своё представление (модель) о том, что повышает (или понижает) вероятность для молекулы быть синтезированной.

1.4.3 Методы моделирования синтетической доступности

Современное понимание о моделировании синтетической доступности состоит в том, что все методы можно условно разделить на три подхода (группы подходов):

1. Ретросинтетический подход, который представляет собой классический и самый надежный способ оценки синтезируемости молекулы. Он определяется возможностью найти последовательность химических реакций, необходимых для получения целевой молекулярной структуры из коммерчески доступных исходных соединений (КДИС). Это, например, можно

охарактеризовать бинарной метрикой: 1 — возможно найти ретросинтетический путь, что означает, что молекулярная структура оценивается как синтетически реализуемая; 0 — невозможно найти ретросинтетический путь, что означает, что молекулярная структура оценивается как синтетически нереализуемая.

2. Подход на основе дескрипторов, который использует молекулярные или реакционные дескрипторы. Молекулярный дескриптор (МД) — это характеристика молекулярной структуры (например, молекулярная масса и количество хиральных атомов углерода) [66], непосредственно или косвенно связанная со сложностью синтеза и, соответственно вероятностью синтезируемости, тогда как реакционные дескрипторы — это характеристика молекулярной реакции или последовательности реакций (например, участие (да/нет) макроциклизации или количество реакционных шагов).

3. Подход на основе данных, который можно разделить на две группы: подходы на основе статистического анализа данных и подходы, основанные на методах машинного обучения (МО). Статистический подход основан на использовании известных химических пространств в качестве основы для получения статистических выводов о синтетической доступности. Подходы, основанные на машинном обучении, обычно представлены методами обучения с учителем. Следовательно, такие подходы требуют размеченного набора данных (например, бинарного набора данных, содержащего легко синтезируемые и трудно синтезируемые молекулярные структуры).

Исходя из предложенной классификации ММСД по группам подходов и их дальнейшего более детального описания, существующие ММСД были сообразно сортированы, и результаты сортировки представлены в таблице 2.

Таблица 2. Классификация методов моделирования синтетической доступности (закрашенные серым цветом ячейки означают принадлежность к классу в соответствующем столбце)

Подходы к моделированию синтетической доступности	**На основе дескрипторов**	**На основе данных**		**На основе ретросинтеза**
Методы моделирования		**На основе статистики**	**На основе МО**	
Bertz et al [64] Whitlock et al [67] Rücker and Rücker [68] Barone and Chanone [65] Kochev et al [69]				
ASKCOS [70] AiZynthFinder [71] Synthia [72] SciFinder [73] Reaxys [74] Konstantinov et al [75]				
Takaoka et al [76]				
SCS [63] FSscore [77] SYBA [78] SMPNN [79]				
RA [80] RSsvm [81] GASA [82] RetroGNN [83] DeepSA [84] DFRscore [85]				
RASA [86] RScore [87]				
RSPred [87]				
SA [43] BR-SAScore [88]				
ReRSA [89]				

1.4.3.1 Методы моделирования синтетической доступности, основанные на ретросинтетическом анализе

Ретросинтетический анализ, теория и методология которого были разработаны Э. Д. Кори во второй половине двадцатого века [90], является классическим и самым точным методом моделирования синтетической доступности. Однако, если он выполняется вручную, то это представляет собой очень трудоемкий процесс. Системы планирования синтеза с помощью компьютера (*Computer-Aided Synthesis Planning,* CASP) рассматриваются как автоматизированные альтернативы ручному ретросинтезу, которые позволяют сэкономить время медицинских химиков. Первые исследования в этой области появились в 1980-1990-х годах [91,92]. Наиболее распространенные открытые CASP-платформы сегодня — это ASKCOS [70] и AiZynthFinder [71,93]. Кроме того, коммерческие платформы, такие как SciFinder [73] и Reaxys [74], предлагают доступ к ретросинтетическим путям на основе извлеченных примеров из известных синтетических случаев. Программное обеспечение Synthia (ранее известное как Chematica) [72] можно считать самым ярким примером автоматизированных ретросинтетических движков. Synthia основана на более чем 100 000 вручную закодированных правил (шаблонов химических реакций) экспертами-химиками [94]. Возможно, это единственный коммерчески доступный инструмент с достойной производительностью, учитывающий хемоселективность, региоселективность и стереоселективность. В то же время, даже в автоматизированном виде, подход к моделированию синтетической доступности на основе ретросинтеза не может рассматриваться как опция для генеративных химических выходов из-за ограниченной скорости алгоритма, поскольку получение ответа для одной молекулярной структуры с помощью Synthia может занять несколько минут. Другой пример: ретросинтетический анализ 200 000 соединений с использованием AiZynthFinder занял 239 дней [80]. Таким образом, CASP-системы неприменимы для генеративных химических движков, которые производят десятки структур в секунду.

В то же время, ключевая проблема методов моделирования синтетической доступности, использующих шаблоны реакций, заключается в проблеме генерализации шаблонов, которая по своему существу связана с базовой проблемой онтологии и эпистемологии — проблемой общего и частного, единого и единичного. По аналогии, в органической химии существует общепринятая концепция именных реакций или именных синтетических методов, при этом таковой реакции необязательно нести имя известного ученого. В качестве примеров последних можно вспомнить пинаколиновую перегруппировку или альдольно-кротоновую конденсацию. Тем не менее, на практике, у каждой такой реакции

есть некоторая относительно ограниченная область применимости, не достигающая границ функциональных классов и/или границ применимости установленного или предполагаемого механизма реакции. В современных научных публикациях по синтетической органической химии принято описывать область применимости (*scope*) метода. Однако принципиально установить границы области применимости метода при помощи современных методов моделирования химических реакций не представляется возможным. В мире органической химии последнее слово всё ещё сохраняется за экспериментальным фактом, и потому любые попытки наперед всеобщим образом и окончательно определить границы применимости синтетического метода обречены на неудачу. Таким образом, любая генерализация метода является весьма условной.

Необходимо отметить, что наша научная группа также занимается разработкой высокопроизводительного CASP-движка [75], который на данный момент является составляющей платформы Chemistry42. Тем не менее, поскольку описание разработки и методологии этого инструмента является крайне объемным, а сам CASP-движок может быть предметом отдельного диссертационного исследования, в настоящей диссертационной работе данный инструмент будет упомянут лишь косвенно в ходе валидации метода моделирования синтетической доступности ReRSA (см. п. 3.2.3.8).

1.4.3.2 Методы моделирования синтетической доступности, основанные на дескрипторах

Исторически первыми моделями синтетической доступности были модели, основанные на концепции молекулярной сложности, которая является своего рода эвристикой СД, базирующейся на молекулярных дескрипторах. Молекулярная сложность рассматривается как модель, выведенная из определенных исчисляемых и интерпретируемых молекулярных свойств и структурных особенностей, непосредственно связанных с синтетическими сложностями. Например, спиро-сочленение может считаться относительно сложным структурным мотивом. Все методы моделирования молекулярной сложности [64,65,67,68,95] являются высокопроизводительными, поскольку требуемые молекулярные дескрипторы могут быть быстро рассчитаны. Однако это также является недостатком этих методов, поскольку сложность молекулярного графа слабо коррелирует с реальной синтетической осуществимостью. Например, некоторые сложные соединения (например, стероидные производные) могут быть легко синтезированы из коммерчески доступных и дешевых исходных материалов, которые с той же точки зрения также являются структурно сложными [62]. Таким образом, молекулярная (структурная) сложность сама по себе не может рассматриваться как адекватная модель СД для таких структур. Тем не менее, нельзя

полностью отрицать возможность ассоциации СД со структурной сложностью, поэтому весьма вероятно, что адекватная модель СД может частично опираться и на структурные дескрипторы, которые будут умело скомбинированы с другими факторами, влияющими на СД, например контекст коммерчески доступных исходных соединений. С другой стороны, дескрипторы, основанные на реакциях, также могут рассматриваться в качестве одного из элементов модели синтетической доступности [86]. Например, длина синтетических путей может быть использована в качестве эвристики для оценки достижимости синтетической кампании.

1.4.3.3 Методы моделирования синтетической доступности, основанные на анализе данных

Расцвет подходов к моделированию СД, основанных на данных, позволил получать ценные выводы из больших наборов данных реальных молекул для более точной оценки синтезируемости новых молекулярных структур. Результаты статистического анализа, полученные из данных, могут быть включены в формулы, соответствующие модели СД, или, в случае методов обучения с учителем, размеченный набор данных может быть использован для обучения модели для решения задач классификации или регрессии. Однако этот подход имеет свои собственные проблемы. Основные проблемы включают:

1. Неполноту охвата химического пространства в наборе данных;
2. Выбор молекулярного представления данных;
3. Маркировку набора данных, если применяются методы обучения с учителем.

Остановимся на каждой из проблем подробнее.

1.4.3.3.1 Проблема охвата химического пространства

Определение химического пространства само по себе является сложной задачей, а определение его синтетически осуществимой части представляет еще большую проблему [96]. Данные можно получить либо из открытых баз данных (например, PubChem, ChEMBL), либо из коммерчески доступных наборов данных (например, запасы поставщиков), однако многие исторические библиотеки фармацевтических компаний и академических исследовательских групп не будут раскрыты в течение неопределенного периода времени, что сохраняет статус *terra incognita* для значительной части химического пространства с точки зрения синтетической доступности [97].

1.4.3.3.2 Проблема выбора молекулярного представления

Выбор молекулярного представления [98] для моделирования СД имеет решающее значение, так как оно должно адекватно отражать синтетическую логику. Следуя этой логике, в идеале молекула должна быть представлена в виде дерева синтонов [90], где синтоны организованы иерархически. Однако структура данных, организованная таким образом, слишком сложна. Чтобы минимизировать сложность представления, применяют несколько подходов:

1. Представление на основе дескрипторов [76,99] включает ограниченный структурный контекст и требует выбора дескрипторов, что само по себе может быть сложной задачей. Если дескрипторы выбираются в режиме без учителя, интерпретируемость такого подхода вызывает сомнения.
2. Представление на основе фрагментов или молекулярных фингерпринтов [43,63,78,81,88,89,100] предоставляет более четкое понимание ретросинтетической природы молекулы, которую необходимо фрагментировать, если такие фрагменты достаточно хорошо приближаются к синтонам. Напротив, если фрагментация не связана с ретросинтетическим расщеплением молекулы (например, при генерации моргановских отпечатков [101]), из анализа набора данных можно получить только статистический контекст, основанный на фрагментах, в то время как синтетический контекст будет утерян.
3. Строковое представление (обычно используются SMILES-строки [102]), которое применяется в архитектурах нейронных сетей типа трансформеров [84].
4. Графовое молекулярное представление [82,83,85,103], хотя и считается самым естественным способом представления молекулярной структуры, их способность использоваться для моделирования синтетической доступности СД не является хорошо доказанной.

Фундаментальная проблема всех приведенных выше представлений заключается в том, что они не позволяют учитывать **гипотезу компактности** (*если объекты находятся в близко друг к другу в рамках некоторого представления, они, вероятно, относятся к одному классу* [104]) при моделировании синтетической доступности для химического пространства. Другими словами, хорошее молекулярное представление должно отражать разделение классов ES (*easy-to-synthesize*, легко синтезируемых) и HS (*hard-to-synthesize*, трудно синтезируемых) соединений. На практике это не работает для классических представлений: две очень похожие структуры могут существенно различаться в синтетической доступности — одна из них может быть легко синтезирована, в то время как другая, отличающаяся лишь одним или парой

атомов, не может быть получена в принципе. Такие «разрывы синтезируемости» (*synthesizability cliffs*) (аналогично «разрывам активности» (*activity cliffs*) [4]) не могут быть разрешены с использованием традиционных методов машинного обучения и молекулярных представлений, что существенно ограничивает применимость таких методов.

Моделирование молекулярной динамики рассматривается как единственный способ преодолеть барьер применимости традиционного анализа КССА (количественное соотношение структура-активность, англ. *quantitative structure-activity relationship, QSAR*) и разрешить разрывы активности [105], которые являются следствием потери информации при уменьшении размерности представления в процессе моделирования связывания лиганда с белком. Аналогично, чтобы не потерять синтетические знания о структуре, необходимо моделировать синтез *in silico*, что можно сделать с помощью ретросинтеза *in silico*. В то время как ретросинтез является самым точным способом оценки синтезируемости молекулы и представляет собой “динамический” процесс, при применении же “статического” представления для молекулярной структуры наблюдается потеря информации. Таким образом, усиление представления ретросинтетическими выводами было бы полезным для более точного моделирования синтетической доступности. В этом смысле теоретически возможно представить набор дескрипторов, которые прямо или косвенно учитывают выводы из ретросинтеза, хотя это потребует значительных вычислительных ресурсов, однако такого представления пока не существует. Другим возможным подходом к учету ретросинтетической механики является внедрение синтетических правил в системы классификации по аналогии с тем, как физические законы интегрируются в сетевые архитектуры при применении так называемого подхода машинного обучения, подкрепленного знаниями о физике явления (*Physics-Informed Machine Learning (PIML)*) [106].

1.4.3.3.3 Проблема разметки набора данных

Разметка обучающего набора данных для “обучения с учителем” является отдельной проблемой. Сообщается о попытках вручную разметить синтетическую доступность (СД) для молекулярных структур [76]. Однако это трудоемкий процесс, который, во-первых, делает размер обучающего набора относительно небольшим, а во-вторых, субъективность экспертных оценок является ключевой проблемой такого подхода [43,99,107]. Другим подходом к маркировке наборов данных для обучения с учителем является использование CASP-движков для бинарной классификации молекулярных структур на легко- (ретросинтетический путь был найден) и трудно-синтезируемые (ретросинтетический путь не был найден) классы [80–82,84]. Безусловно, предложенные методы, использующие для

обучения моделей выходные данные CASP-инструментов, способны быстро обрабатывать огромное количество структур, и в соответствующих исследованиях были получены хорошие коэффициенты корреляции на тестовых выборках, которые обычно также являются продуктами соответствующих CASP-движков. Однако эти методы характеризуются несколькими типичными проблемами. Во-первых, обучающий набор наследует все ошибки и недостатки CASP-инструментов. Поскольку существует очень мало CASP-движков, полностью учитывающих проблемы хемоселективности, региоселективности и стереоселективности, результаты работы посредственного CASP не могут быть по-настоящему интерпретируемые как “легко-синтезируемые” или “трудно-синтезируемые”, так как сами генерируемые пути могут быть ненадежными. Во-вторых, если пути генерируются в большом количестве, отсутствие ручной проверки приводит исследователей, использующих подходы на основе результатов CASP, к распространенной проблеме анализа данных: “мусор на входе, мусор на выходе”.

1.4.3.3.4 Методы моделирования синтетической доступности, основанные на статистическом анализе референсного химического пространства

Даже несмотря на то, что классификация подходов по моделированию синтетической доступности, основанных на данных, представлена следующим образом:

- Подходы на основе статистического анализа референсного химического пространства;
- Подходы, основанные на машинном обучении;

мы считаем, что на сегодняшний день не существует чисто статистических методов моделирования синтетической доступности (СД), и модели СД, использующие статистический анализ референсного химического пространства, на самом деле используют смешанные методы, такие как, например, комбинация подходов на основе дескрипторов и статистики [43].

Все статистические подходы основаны на гипотезе о том, что молекулярная структура, содержащая только общие структурные мотивы, встречающиеся в коммерчески доступных лекарственных веществах (или в синтезированном и зарегистрированном химическом пространстве), вероятно, будет легко синтезируемой. Таким образом, частота встречаемости структурных мотивов в репрезентативном химическом пространстве составляет меру синтетической доступности для молекулярных структур, содержащих эти структурные мотивы. Чем чаще встречается структурный мотив в репрезентативном химическом

пространстве, тем более синтетически доступным он делает любую родительскую молекулярную структуру.

Наиболее известным и широко используемым алгоритмом является **SA Score** (*Synthetic Accessibility Score*) [43], исходный код которого открыт и встроен в библиотеку RDKit [27]. SA Score представляет собой смешанный подход, основанный на молекулярных дескрипторах и статистических данных из исторических синтетических знаний. Эти данные получены из анализа общих структурных особенностей молекулярных фрагментов (в данном методе фрагментом считается подструктура молекулы, полученная путем фрагментации молекулы с использованием метода генерации молекулярных фингерпринтов ECFC_4) в подготовленной базе данных уже синтезированных молекул (подмножество PubChem [108]). Вторая часть SA Score основана на рассчитанном штрафе, который характеризует наличие сложных структурных особенностей в молекулах, таких как количество хиральных центров, количество спиро-точек, наличие макроциклов и т. д. В результате SA Score при помощи вышеописанной гипотезы демонстрирует компромисс между быстрым подходом, основанным на структурной сложности (структурных дескрипторах), и ресурсозатратными полномасштабными подходами на основе ретросинтеза.

1.4.3.3.5 Методы моделирования синтетической доступности, основанные на машинном обучении

Среди моделей синтетической доступности (СД), основанных на машинном обучении (МО), SCScore [63] и SYBA [78] широко используются в качестве эталонов для сравнения. В то же время стоит упомянуть RA Score [80], поскольку он использует подход, отличающийся от чисто МО-ориентированных методов моделирования СД, таких как SCScore и SYBA.

Упомянутый выше **SCScore** (*Synthetic Complexity Score*) [63] является наглядным примером моделей синтетической доступности, основанных на данных, которые используют прецедентные знания о химических реакциях для обучения функции аппроксимации для оценки синтетизируемости молекулярных структур. В качестве функции аппроксимации SCScore использует полносвязную искусственную нейронную сеть, которая обучается с использованием стандартного алгоритма обратного распространения ошибки на большой базе данных известных синтезированных молекул, обладающих свойствами, схожими с лекарственными средствами, и их известных синтетических путей. Ключевая идея SCScore заключается в том, чтобы обучить такую ранжирующую функцию оценивающую **синтетическую сложность** (важно не путать с синтетической доступностью или синтетической осуществимостью), которая для продуктов реакции должна быть больше, чем

для любых отдельных реагентов в этой реакции. Однако это локальное предположение верно только для крайних случаев (начало синтеза и конец синтеза), когда оно применяется для последовательности реакций. В то время как основная проблематика ретросинтетического анализа кроется в середине синтетических последовательностей, где различие между «продуктами как более сложными» и «реагентами как менее сложными» исчезает. Более того, это на первый взгляд интуитивное предположение противоречит самой природе синтеза, когда некоторые стратегические реакции (например, защиты) применяются для снижения синтетической сложности с точки зрения SCScore, временно делая защищенное соединение более пригодным для использования в качестве реагента из-за улучшенной функциональности (установка защитной группы решает проблемы селективности). Противоречие заключается в том, что правильная метрика синтетической сложности должна монотонно уменьшаться от целевой молекулы к исходным соединениям. В качестве эксперимента авторы статьи предлагают рассмотреть последовательности реакций, которые ведут к синтезу малых лекарственных молекул, одобренных национальными регуляторами в 2015 году [109]. Авторы делают предположение, что предлагаемая ими метрика синтетической сложности способна коррелировать с числом стадий необходимым для синтеза молекул. Однако сама базовая гипотеза SCScore выглядит несостоятельной, поскольку органический синтез постоянно эксплуатирует реакции, снижающие сложность молекулы, например реакции удаления защитной группы, ретро-реакции, реакции элиминирования. Напротив, установка защитной группы или нескольких ортогональных защитных групп может сделать молекулу более функциональной и простой в плане дальнейшей стратегии органического синтеза, и она скорее будет востребована в качестве реагента, несмотря на свою более высокую сложность, чем незащищенный исходник. Таким образом реакция постановки защитной группы должна приводить к понижению синтетической сложности, то есть SCScore защищенного продукта должна быть ниже, чем SCScore исходного реагента, что противоречит базовой гипотезе метода и приводит к принципиальной невозможности использовать SCScore для навигации по ретросинтетическому древу. Подобная неопределенность и несвязность между положением молекулы в глобальном пространстве реакций и локальным участием в конкретных синтезах авторами статьи не комментируется. Тем более непонятно, как пользоваться SCScore для отбора молекул на синтез. Таким образом положение молекулы в отдельной реакции (в качестве продукта или реактанта) или в цепочке реакций совершенно не обязательно связано с ее синтетической сложностью. Так, в статье не проведен анализ вклада молекул с подобной двоякой природой и насколько они влияют на точность предсказания синтетической сложности.

Synthetic Bayesian Accessibility (SYBA) [78] — это модель синтетической доступности на основе фрагментов, предназначенная для быстрого классифицирования органических соединений как легко- (ES, *easy-to-synthesize*) или трудно-синтезируемых (HS, *hard-to-synthesize*). Она основана на наивном байесовском классификаторе, который используется для присвоения вкладов в оценку SYBA отдельным фрагментам на основе их частоты в базе данных ES и HS молекул. Наиболее спорным аспектом модели SYBA является то, что выбранные для обучения HS молекулы представляют собой крайний процентиль явно недоступного химического пространства (см. рис 6), в то время как недоступное пространство значительно больше.

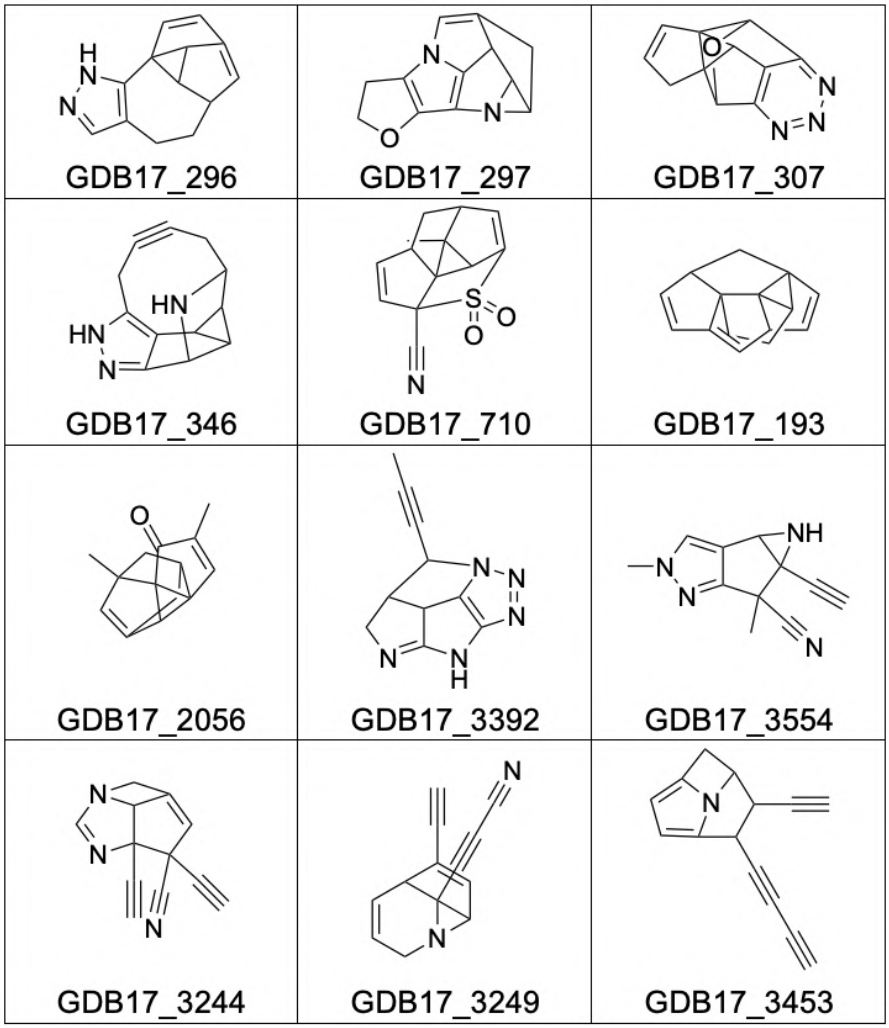


Рисунок 6. Примеры трудно-синтезируемых молекулярных структур из обучающей выборки метода моделирования SYBA.

Более того, самая интересная часть недоступного химического пространства выглядит таким образом, что среднестатистический химик-синтетик должен колебаться, прежде чем немедленно пометить его как недоступное. Такие химические структуры составляют значительную часть предсказанных ложноположительных результатов.

Retrosynthetic Accessibility score (RAscore) [80] представляет собой классификатор на основе МО для моделирования синтетической доступности, обученный на результатах, сгенерированных с помощью инструмента планирования синтеза AiZynthFinder [110]. Таким образом, этот подход можно формально считать смешанным, поскольку он сочетает в себе ретросинтетический подход и подход, основанный на данных (подход, основанный на машинном обучении). В рамках метода использовались следующие алгоритмы классификации:

- классификатор на основе прямого распространения нейронной сети,
- классификатор XGBoost,
- классификатор случайного леса.

Для каждого алгоритма использовались расширенные молекулярные отпечатки соединений с учетом связи (*extended connectivity fingerprints*, ECFP6) размерностью 2048 и радиусом, равным 3, а также подсчеты ECFP6 с признаками, сгенерированными с помощью RDKit. Основная концептуальная проблема такого подхода заключается в том, что маркировка синтетической доступности на основе AiZynthFinder вызывает сомнения, так как этот инструмент планирования синтеза на момент разработки RAscore не предоставлял высококачественные пути, учитывающие проблемы селективности.

В качестве наглядного примера рассмотрим ретросинтетические схемы двух органических структур, предсказанные AiZynthFinder (см. рис. 7) и приведенные в оригинальной статье [80]. В предложенной синтетической схеме вещества **A1** для синтеза используется тиазолоиндол **A8**, который не является коммерчески доступным соединением, таким образом, для **A8** требуется дополнительное рассмотрение методов его синтеза. На стадии **A5→A4** предлагается обработка сульфокислоты тионилхлоридом для получения сульфонилхлорида, при этом в результате реакции выделяется соляная кислота, способная гидролизовать имин, помимо этого, получаемый сульфонихлорид способен вступать в реакцию с незащищенным индольным атомом азота и иминной группировкой. На стадии **A3→A2** предлагается гидролиз в водной среде, что также не селективно в присутствии имина. На стадии **A2→A1** проводится дегалогенирование. Условие для селективности этой реакции, возможно, можно и подобрать, тем не менее, возникает вопрос, с какой целью нужно было вводить бром. В предложенной схеме синтеза вещества **B1** использованы коммерчески доступные стартовые материалы, однако вопросы вызывают стадия **B5→B4**, подразумевающая восстановление нитрогруппы в присутствии в молекуле имина; стадия замыкания пятичленного цикла аминобензотиазола **B3→B2**, которая в данном случае является не

региоселективной; а также стадия восстановительного деаминирования **B2→B1**, которая не селективна в присутствии иминной группировки.

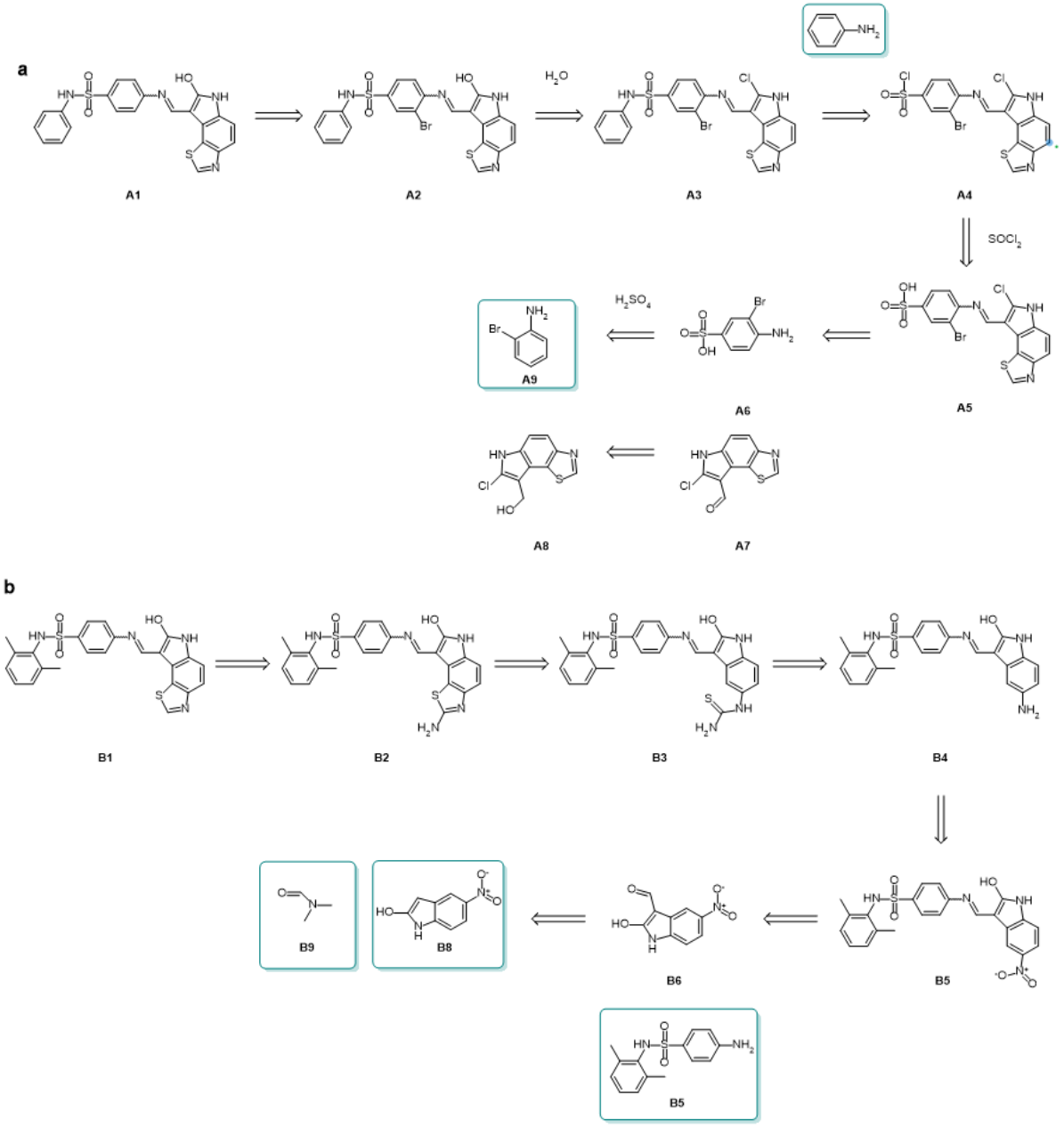


Рисунок 7. Ретросинтетические схемы соединений **A1** (**a**) и **B1** (**b**), предсказанные CASP-инструментом AiZynthFinder. Зеленой рамкой выделены коммерчески доступные стартовые материалы.

2. Материалы и методы

2.1 Метод моделирования структурных трендов MCE-18

Поскольку ранее (см. раздел 1.3) были обозначены объективные причины о наличии структурных трендов и структурной эволюции, а также диалектическое отношение между структурной новизной и синтетической доступностью молекулярных структур, производимых алгоритмами генеративной химии (см. раздел 1.2.4), было предложено создать новый алгоритм, позволяющий оценивать соответствие молекулярных структур современным трендам, наблюдаемым среди структур, запатентованных крупнейшими фармацевтическими компаниями. Новый алгоритм предложено назвать MCE-18 (*Medicinal Chemistry Evolution by 2018*, эволюция медицинской химии по статусу на 2018 год), исходя из эволюционного восприятия картины, изменяющихся с годами структур потенциальных и зарегистрированных лекарственных веществ.

2.1.1. Сбор баз данных для метода MCE-18

2.1.1.1 База данных молекулярных структур из фармацевтических патентов

Чтобы избежать реагентов и промежуточных соединений, мы извлекли только уникальные структуры из фармацевтических патентов крупнейших фармацевтических компаний, доступных в базе данных Clarivate Analytics Integrity в разделе “Patents” [111]. Все эти молекулы были заявлены составителями базы данных как соединения-лидеры в рамках каждого из патента и не были найдены среди реагентов или промежуточных соединений. Каждой структуре был присвоен самый ранний год (дата приоритета), в который она появилась. Для этого анализа были выбраны 23 ведущие фармацевтические компании. Были загружены патентные записи о новых лекарственных веществах от этих компаний с датами приоритета с 1950 по 2018 год последовательно год за годом, обеспечивая фильтрацию дубликатов более старых лекарственных веществ, если таковые попадались в более поздний период. Процедура фильтрации для изоляции только уникальных записей дала в итоге 30 153 записи о молекулярных структурах. Затем отчеты были повторно загружены в виде файлов структурных данных (sdf) из базы данных Clarivate Analytics Integrity “Drugs & Biologics” и предварительно обработаны с помощью программного обеспечения ChemoSoft. Таким образом, каждому соединению было присвоено поле “год патента”, и все элементы были объединены в один sdf-файл, который затем подвергся первичной подготовке: структуры были нормализованы, нейтрализованы, противоионы были удалены, а ошибки в структурах были исправлены вручную. К базе данных были применены мягкие медицинские химические

фильтры, что привело к исключению металлоорганических, кремнийорганических и фосфорорганических соединений, изотопов и т. д. На первом этапе предварительной обработки было исключено 957 структур. Затем был удален 281 образец из-за несоответствия между "годом патента" и датой запуска: лекарство было запущено до записи патента компанией, не входящей в исследуемую группу. Также были исключены структуры, отвечающие следующим критериям: молекулярная масса (MW) > 1200, более 20 атомов кислорода, более 10 атомов фтора, а также структуры без углеродных атомов. Дубликаты были удалены, и база данных была сокращена до 28 161 образца. Чтобы увеличить разнообразие соединений и уменьшить количество чрезмерно представленных хемотипов, в ChemoSoft была проведена процедура кластеризации. В качестве меры 2D-сходства использовалась обычная метрика Танимото. Соединения с оценками сходства выше 0.5 были отнесены к одному кластеру (минимум 10 записей на кластер). Затем структуры в каждом кластере были ранжированы по их коэффициентам разнообразия, и образцы с самыми высокими 10% (для кластеров, содержащих менее 200 записей) или 5% (для кластеров, содержащих более 200 записей) оценок разнообразия, а также все соединения, не отнесенные к кластеру, были сохранены. Итоговая база данных содержала 24 232 структуры. В результате предварительной обработки и последующей нормализации было исключено в общей сложности ~4 000 структур (14%). Исключенные структуры могли иметь некоторое значение, однако было очень важно выполнить перечисленные операции для достижения адекватного статистического результата и обеспечения надежного *in silico* моделирования. Например, для извлечения реальных структурных трендов было необходимо путем описанной кластеризации “подавить” имевшиеся в изобилии и создававшие статистический шум “старые” хемотипы: морфины, пенициллины, цефалоспорины, стероиды, макролиды, тетрациклины и фторхинолоны.

2.1.1.2 База данных одобренных лекарственных веществ

Исходная коллекция из 4 385 зарегистрированных лекарственных препаратов была получена из базы данных Clarivate Analytics Integrity “Drugs & Biologics”. Мы обнаружили, что информация о дате выпуска отсутствовала у 1 312 препаратов, поэтому эти записи были исключены. Оставшиеся 3 073 записи были повторно загружены из “Drugs & Biologics” в формате sdf. Из оставшихся записей 1 383 не содержали никакой информации о химической структуре (биопрепараты), поэтому они также были исключены. К каждому образцу была присвоена самая ранняя дата одобрения/регистрации/запуска. В результате 1 690 записей с их основной датой запуска были включены в окончательный набор данных без дубликатов. Затем база данных была предварительно обработана, как описано выше в 2.1.1.1, что привело к исключению еще 14 записей. Затем была проведена кластеризация на основе коэффициента

Танимото по аналогии с описанной в 2.1.1.1 процедурой. Молекулы с высокой структурной схожестью были удалены из каждого кластера (минимальное количество образцов в кластере составляло 10). Все соединения, не отнесенные к кластеру, были сохранены. Предварительная обработка и нормализация дали 1 370 уникальных структур.

2.1.1.3 База данных лекарственных веществ на разных этапах разработки

Тот же источник данных был использован для сравнения лекарственных веществ, прошедших разный путь от соединений-лидеров до зарегистрированных лекарственных веществ. Таким образом различались молекулы, которые:

1. использовались в качестве соединений-лидеров,
2. проходили клиническую оценку как кандидаты в лекарственные препараты (фазы клинических испытаний I–III),
3. были зарегистрированы как лекарственные препараты.

При этом рассматривалась только самая высокая достигнутая контрольная точка для каждой из молекул. Для каждой категории была проведена предварительная обработка и нормализация в соответствии с описанной выше (см. разд. 2.1.1.1 и 2.1.1.2) процедурой. В результате окончательная база данных содержала около 30 000 соединений лидеров, 1 678 клинических кандидатов в фазе I, 1 837 клинических кандидатов в фазе II, 464 клинических кандидата в фазе III и 1 370 зарегистрированных лекарственных веществ.

2.1.2 Молекулярные дескрипторы

Для всех соединений в базе данных были рассчитаны ключевые молекулярные дескрипторы с использованием программного обеспечения ChemoSoft [112] и SmartMining [113]. Эти дескрипторы включают MCE-18 (см. разд. 2.1.3), молекулярную массу (MW), LogP (липофильность, рассчитанный коэффициент распределения в системе 1-октанол/вода), $LogS_w$ (предсказанная растворимость в воде), PSA (площадь полярной поверхности, Å²), HBD (число потенциальных доноров водородных связей), HBA (число потенциальных акцепторов водородных связей), а также SS (общий электро-топологический индекс) и такие показатели, как AR, NAR, CHIRAL, SPIRO, NCSPTR, Q^1 и Fsp^3, которые определены ниже. Для выбранных дескрипторов были рассчитаны значения парного t-критерия Стьюдента). Мы использовали значения MCE-18, MW, PSA, HBD, HBA, $LogS_w$ и SS для *in silico* моделирования, и этот набор параметров был выбран в соответствии с их теоретическим влиянием на изучаемое явление.

2.1.3 Функция MCE-18

На основании вышеупомянутых вопросов мы исследовали основные тенденции в медицинской химии и разработке лекарств, сосредоточив внимание на эволюции химических структур лекарств, которые исследовались в рамках клинических испытаний, были выведены на фармацевтический рынок или заявлены в патентных записях. Мы предлагаем функцию MCE-18 в качестве нового молекулярного дескриптора, который может эффективно оценивать структуры по их новизне и текущему потенциалу соединений-лидеров в отличие от простого и во многих случаях ложноположительного индекса Fsp^3, и он определяется следующей формулой (1):

$$\text{MCE-18} = \left(\text{AR} + \text{NAR} + \text{CHIRAL} + \text{SPIRO} + \frac{\overbrace{\text{sp}^3 + \text{Cyc} - \text{Acyc}}^{\text{NCSPTR}}}{1 + \text{sp}^3} \right) \times Q^1, \qquad (1)$$

где AR обозначает наличие ароматического или гетероароматического кольца (0 или 1), NAR обозначает наличие алифатического или гетероалифатического кольца (0 или 1), CHIRAL обозначает наличие хирального центра (0 или 1), SPIRO обозначает наличие спиро-точки (0 или 1), sp^3 обозначает долю sp^3-гибридизованных углеродных атомов (от 0 до 1), Cyc обозначает долю циклических углеродных атомов, которые являются sp^3-гибридизованными (от 0 до 1), Acyc обозначает долю ациклических углеродных атомов, которые являются sp^3-гибридизованными (от 0 до 1), и Q^1 обозначает нормализованный квадратичный индекс [66]. Последний член в скобке представляет собой взвешенную долю sp^3-гибридизированных атомов углерода, находящихся в циклах (NCSPTR, *non-cyclic sp^3-atom ratio*), значение которого подробно обсуждается в разд. 3.3.2.

2.2 Метод моделирования синтетической доступности ReRSA

Ввиду вышеуказанных недостатков (см. разд. 1.4.3) существующих методов моделирования синтетической доступности (ММСД), в целях эффективной оценки синтезируемости молекулярных структур мы предлагаем новый ММСД, который комбинирует все преимущества существующих ММСД: как учет фактически синтезированного пространства в случае подхода на основе анализа данных, так и учёт ретросинтетического фактора, как наиважнейшего для наиболее адекватной оценки синтезируемости, а также фактор структурных дескрипторов, ассоциированных с понятием структурной сложности. Предлагаемый ММСД носит название **ReRSA** (***Re****trosynthesis-****R****elated*

***S**ynthetic **A**ccessibility*, связанная с ретросинтезом синтетическая доступность) ввиду того, что основной алгоритм метода, в сущности, представляет собой подобие ретросинтетического анализа. Но поскольку полноценный ретросинтетический анализ в методе не используется по описанным ранее причинам (см. разд. 1.4.3.1), сам метод нельзя трактовать как *“основанную на ретросинтезе”* (*retrosynthesis-based*, англ.) модель синтетической доступности, а лишь как *“связанную с ретросинтезом”*. Аналогично, по той же причине, мы не называем саму процедуру фрагментации на синтоноподобные фрагменты (см. разд. 2.2.3) ретросинтетическим анализом, а лишь квази-ретросинтетическим анализом или квази-ретросинтетической фрагментацией.

2.2.1 Базовая гипотеза метода и терминология

Гипотеза о том, что если фрагменты ранее синтезированных молекул будут находиться в новых молекулярных структурах, то это повышает вероятность синтезируемости последней, является базовой для метода моделирования синтетической доступности SA Score [43]. Более того, чем чаще в референсном датасете ранее синтезированных молекул будет находиться некоторая подструктура, то тем выше будет её положительный вклад в синтетическую доступность содержащих эту подструктуру молекулярных структур. На наш взгляд эта гипотеза является логичной и за основу нового метода моделирования СД мы брали именно её.

Тем не менее, метод фрагментации ECFC_4 референсного датасета, которым оперирует SA Score, представлялся несовершенным и, более того, не соответствующим задачам моделирования синтетической доступности, поскольку получаемые фрагменты синтетически бессмысленны, что продемонстрировано на рис. 8.

Ранее мы выяснили, что наиболее предпочтительным методом моделирования синтетической доступности является ретросинтетический анализ, как наиболее приближенный к самому процессу синтеза. В ходе ретросинтетического анализа химики оперируют синтонами, которые напоминают фрагменты молекулы. Исходя из вышесказанного представляется, что наиболее удачным методом фрагментации является именно разбиение молекулярной структуры на синтоны. В то же время, проведение полноценного ретросинтетического анализа невозможно для нужд генеративной химии из-за низкой скорости выполнения даже в автоматизированном режиме, поэтому необходимо было найти компромиссный вариант, который бы учитывал желание видеть в фрагментах синтоноподобные сущности и высокую производительность ММСД.

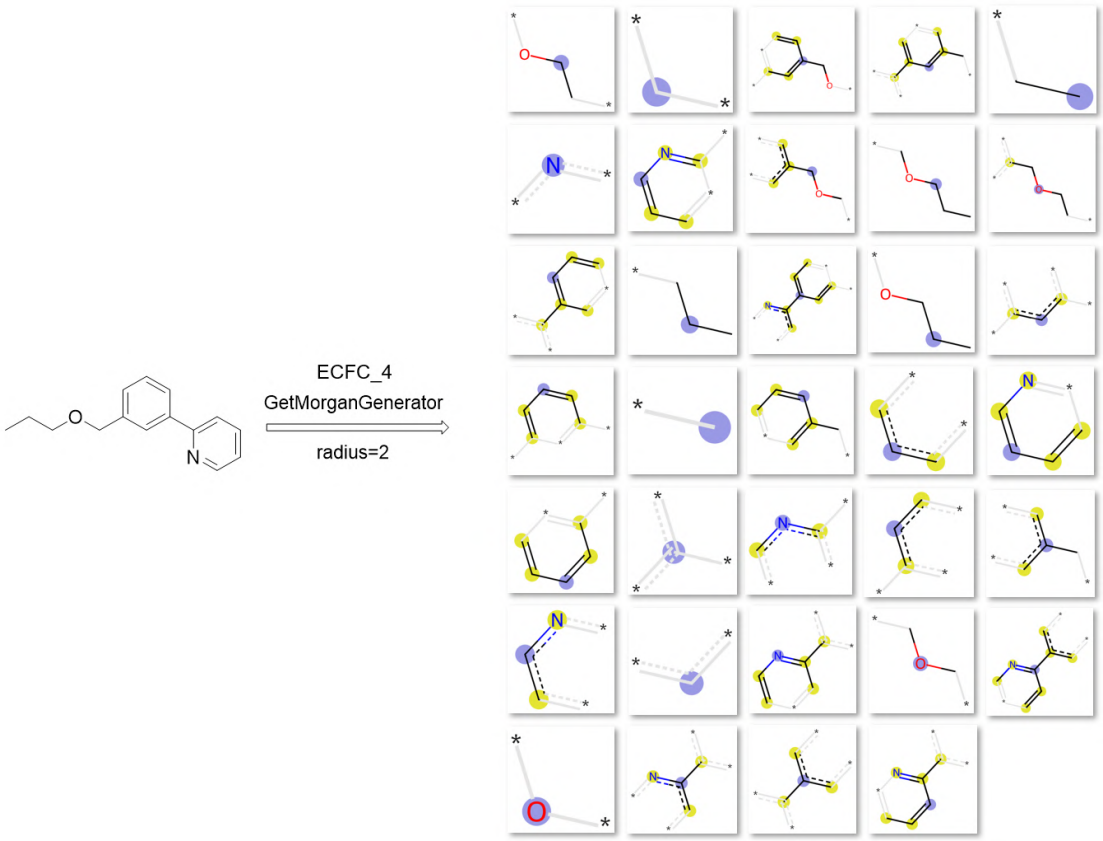


Рисунок 8. Получаемые при помощи функции GetMorganGenerator из библиотеки RDKit. ECFC_4 фрагменты не похожи на синтоны и содержат части произвольным образом разорванных циклов.

После проведения фрагментации синтоноподобные фрагменты будут сконвертированы в соответствующие им синтетические эквиваленты, если оперировать терминами теории ретросинтетического анализа. Последние в свою очередь будут проверены на их наличие в базе данных коммерчески доступных исходных соединений (КДИС).

Параллельно сама исходная молекулярная структура может быть обсчитана при помощи молекулярных дескрипторов, ассоциированных со структурной сложностью. Последняя в свою очередь ассоциируется с низкой вероятностью синтетической осуществимости, хотя это и не всегда является правдой. По этой причине вклад таких молекулярных дескрипторов не должен играть ведущую роль в агрегированной оценке синтетической осуществимости.

Итак, суммируя вышесказанное метод ReRSA представляет собой метод моделирования синтетической доступности молекулярных структур, включающий в себя следующие ключевые операции:

1. Фрагментация молекулярной структуры на синтоноподобные фрагменты и сравнение этих фрагментов с референсным набором фрагментов;
2. Конвертация синтоноподобных фрагментов, полученных на предыдущем этапе, в стартовые материалы и сравнение полученных стартовых материалов с референсным набором стартовых материалов;
3. Расчёт молекулярных дескрипторов, описывающих структурную (молекулярную) сложность, для молекулярной структуры;
4. Агрегация результатов операций (1), (2) и (3) в единую оценку синтезируемости молекулы.

Принципиальная схема метода ReRSA представлена на рис. 9.

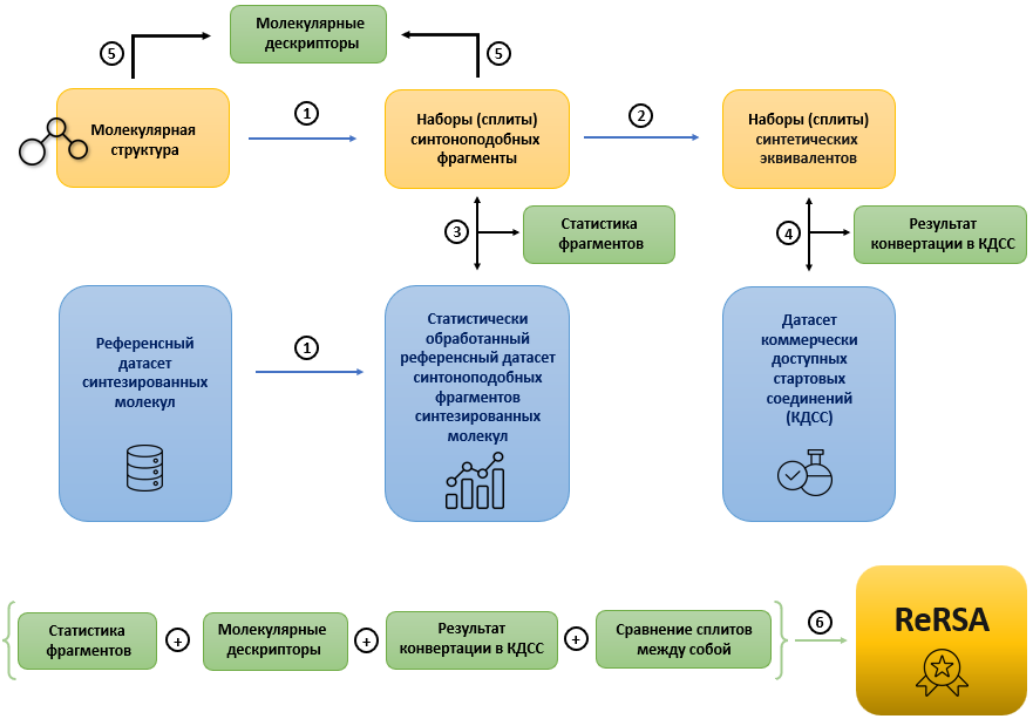


Рисунок 9. Принципиальная схема ММСД ReRSA.

2.2.2 Алгоритм фрагментации

В качестве технического решения, которое бы могло быстро проводить фрагментацию по образу и подобию ретросинтетического анализа, было предложено использовать фрагментацию из ранее упомянутого (см. разд. 1.2.2) генерационного алгоритма BRICS. Действительно, алгоритм разбиения молекулярных структур, реализованный в BRICS напоминает ретросинтетический анализ с той разницей, что разбиение по “стратегическим связям” происходит комбинаторно, то есть разбиваются одновременно все такие связи, а получаемый на выходе набор фрагментов содержит перемешанные результаты всех возможных веток общего виртуального ретросинтетического древа (см. рис. 10).

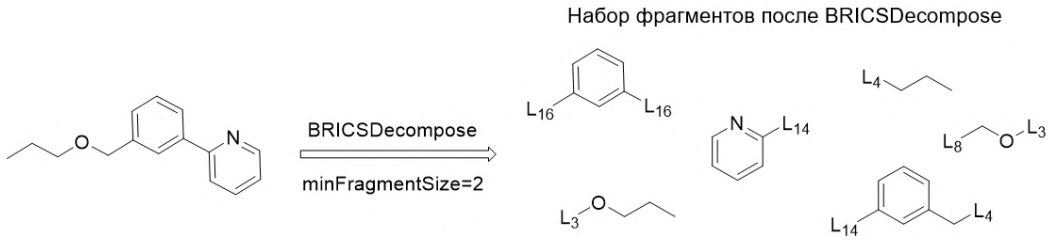


Рисунок 10. Результат квази-ретросинтетической фрагментации 2-(3-пропоксиметил)фенил)пиридина при помощи метода BRICSDecompose.

Тем не менее, даже такая квази-ретросинтетическая фрагментация имеет больше синтетического смысла, чем фрагментация при помощи ECFC_4 фрагментов, более известных как моргановские отпечатки с радиусом 2. В то же время в ходе визуального анализа фрагментов, получаемых после применения метода BRICSDecompose, можно проследить отдельные ветви ретросинтетического древа, приводящие к разным наборам синтоноподобных фрагментов (см рис. 11).

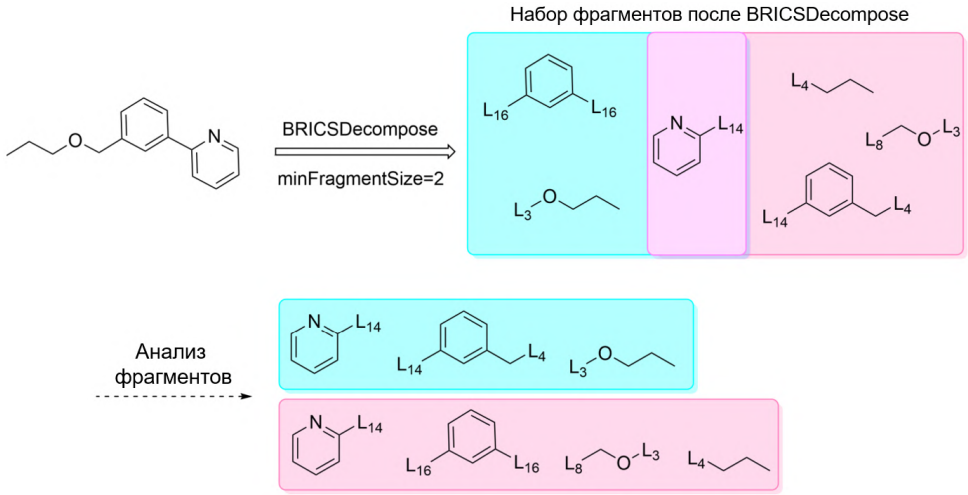


Рисунок 11. Результат анализа набора фрагментов по отдельным веткам ретросинтетического древа, имплицитно генерируемого в ходе работы метода BRICSDecompose.

Таким образом, если модифицировать алгоритм BRICSDecompose, чтобы он агрегировал исходы отличных ветвей ретросинтетического древа, то можно добиться результата работы, который более приближенным образом соответствует классическому ретросинтетическому анализу. Именно эта ключевая модификация базового метода BRICSDecompose была реализована в рамках метода ReRSA.

2.2.4 Робастные реакции для квази-ретросинтетической фрагментации

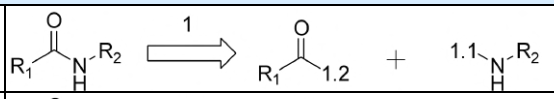
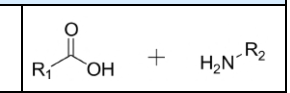
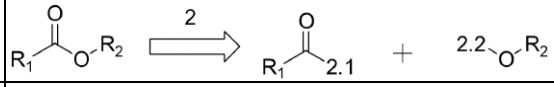
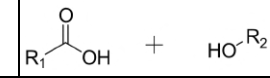
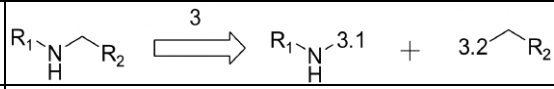
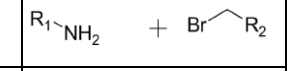
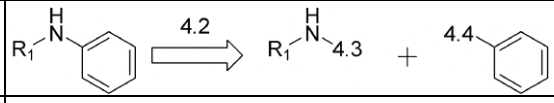
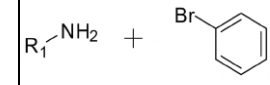
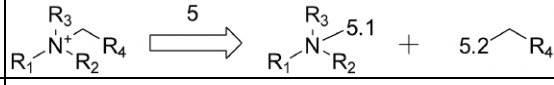
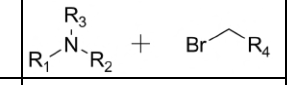
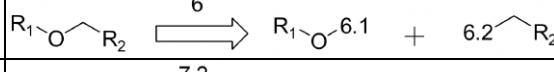
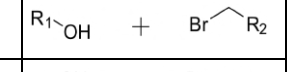
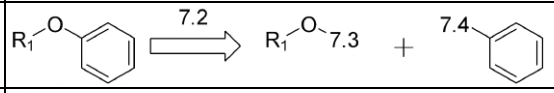
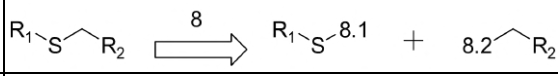
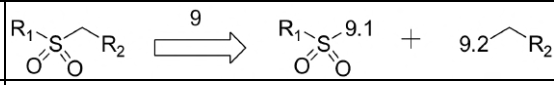
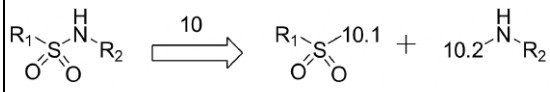
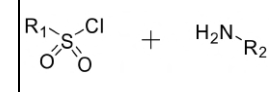
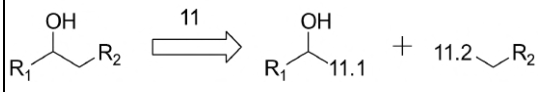
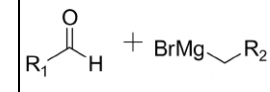
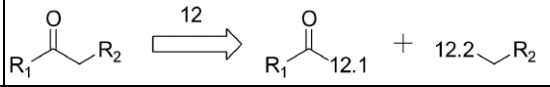
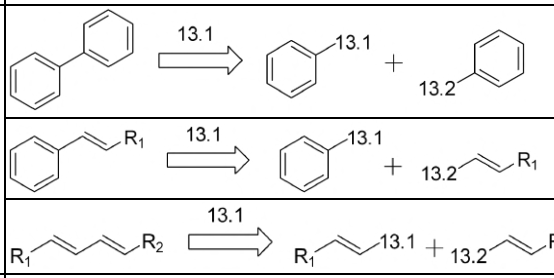
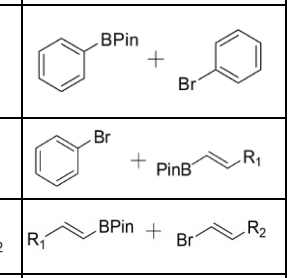
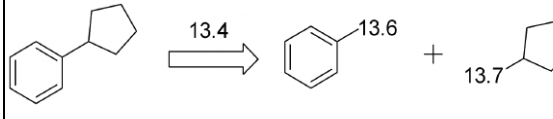
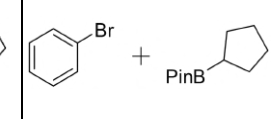
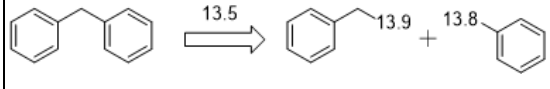
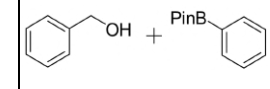
Важнейшей составляющей алгоритма ReRSA, привносящей ретросинтетическую логику в процесс моделирования синтетической доступности, являются ретро-реакции, с

помощью которых, можно осуществлять квази-ретросинтетическую фрагментацию и иные трансформации молекулярных структур. В базовой имплементации алгоритма BRICSDecompose, доступной в RDKit, описывается лишь ограниченное число таких робастных реакций, которыми можно было бы покрыть большую часть молекулярных структур, обладающих подобием к лекарствам. Так, например, в исходном коде BRICSDecompose отсутствует разбиение по сульфонамидной связи и разбиение связи Csp^2-Csp, соответствующее реакции Соногаширы.

С целью снабдить метод ReRSA ретро-реакциями, имеющими значение для медицинской химии, был проведен анализ профильной литературы. Приоритет отдавался тем источникам, которые были посвящены анализу патентов крупных фармацевтических компаний [114,115] и непосредственно обзорам по синтезам зарегистрированных лекарственных веществ, которые публиковались в *Journal of Medicinal Chemistry.* В ходе анализа были приоритезированы 52 ретро-реакции, которые покрывают большую часть реакций, необходимых для синтеза зарегистрированных лекарственных веществ. Описание отобранных робастных ретро-реакций представлено в табл. 3.

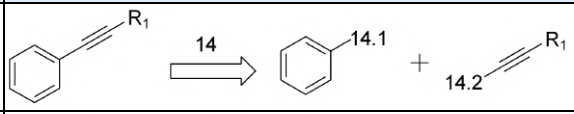
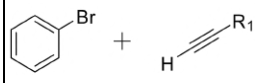
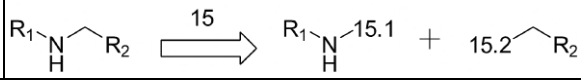
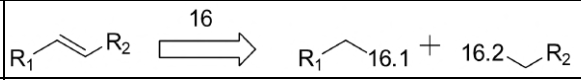
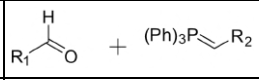
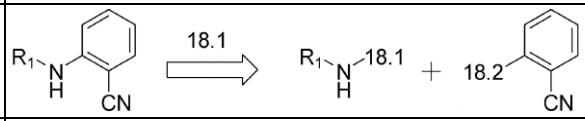
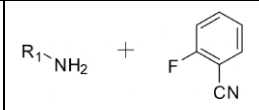
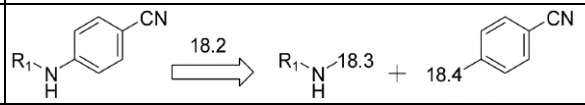
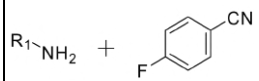
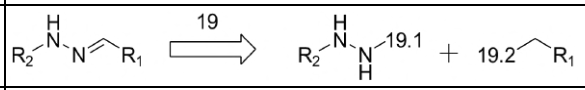
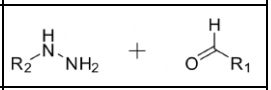
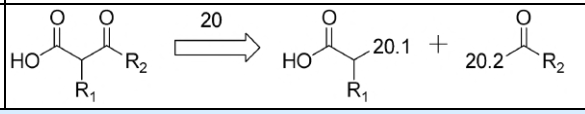
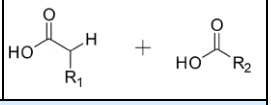
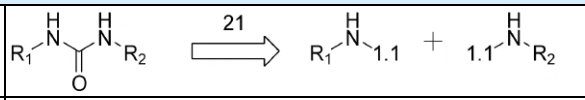
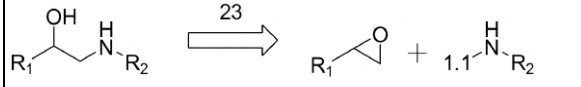
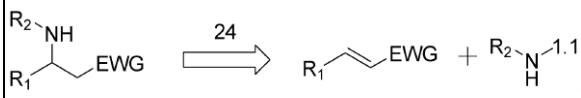
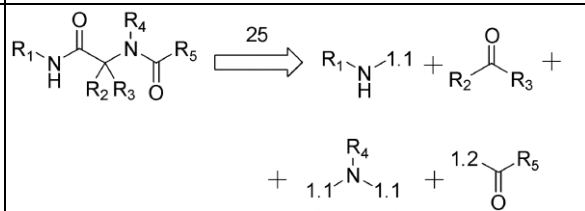
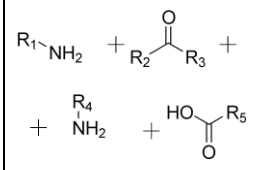
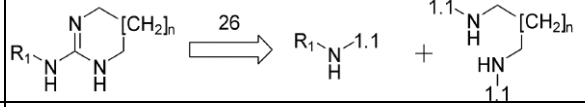
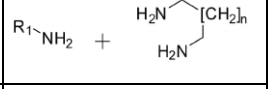
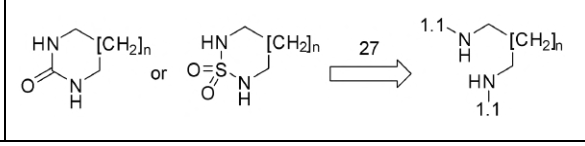
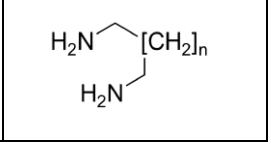
Серьезным ограничением метода BRICSDecompose является отсутствие возможности производить ретросинтетическое разбиение одновременно более чем одной связи. Так, например, ни одна реакция циклизации не может быть корректно осуществлена при помощи метода BRICSDecompose. В связи с этим помимо BRICSDecompose метод ReRSA использует реакции (см. табл. 3, реакции R21-R37), которые выполняются при помощи стандартного метода RunReactants модуля обработки реакции из библиотеки RDKit, причем SMARTS-шаблоны для этих реакций написаны таким образом, чтобы получались фрагменты подобные тем, что производятся при работе BRICSDecompose для корректного учета статистик получаемых синтоноподобных фрагментов. Дополнительный блок реакций, которые не производят фрагменты, но проводят начальную модификацию исходной молекулярной структуры, так же обрабатываются при помощи метода RunReactants. К их числу мы относим, так называемые, однокомпонентные реакции (поскольку на входе и на выходе из реакции используется лишь одна молекулярная сущность), перечисленные в табл. 3. (см. реакции R38-R42). Наконец, особняком стоят реакции макроциклизации, которые применяются только в том случае, есть в молекуле есть макроцикл размером более 7 атомов. Реакции макроциклизации могут приводить как к фрагментам (реакции R45, R47-R52), так и полноценным молекулярным структурам в случае цикл-замыкающего метатезиса (реакции R43, R44 и R46).

Таблица 3. Ретро-реакции, используемые при квази-ретросинтетической фрагментации в алгоритме ReRSA

ID	Трансформ	Схема ретро-трансформации	Пример реактантов
BRICS-подобные реакции			
R1	Ацилирование NH-нуклеофилов		
R2	Ацилирование O-нуклеофилов		
R3	Алкилирование NH-нуклеофилов		
R4.2	Арилирование N-нуклеофилов		
R5	Алкилирование третичных аминов		
R6	Алкилирование O-нуклеофилов		
R7.2	Арилирование O-нуклеофилов		
R8	Алкилирование S-нуклеофилов		
R9	Алкилирование сульфинатов		
R10	Сульфо-ацилирование NH-нуклеофилов		
R11	Реакция Гриньяра с альдегидами и кетонами		
R12	Ацилирование реактивов Гриньяра		
R13.1	Реакция Сузуки-Мияуры между C(sp^2) субстратами		
R13.4	Реакция Сузуки-Мияуры между C(sp^2) и циклическим C(sp^3) субстратами		
R13.5	Реакция Сузуки-Мияуры между C(sp^2) и бензильным субстратами		

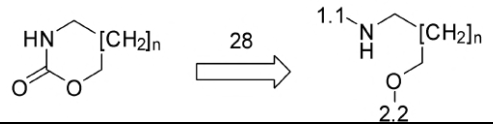
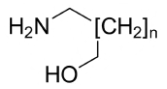
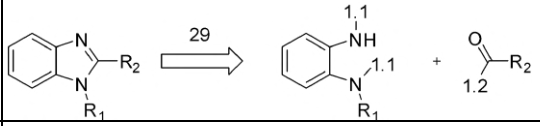
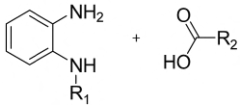
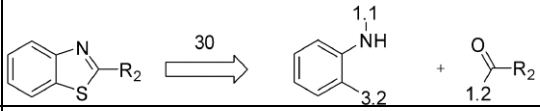
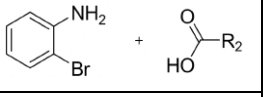
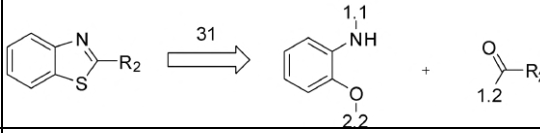
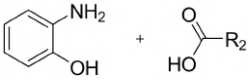
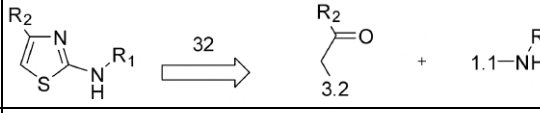
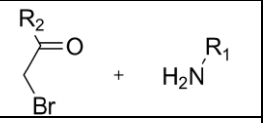
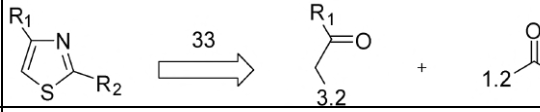
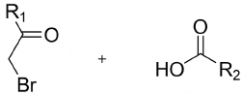
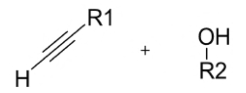
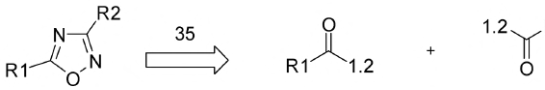
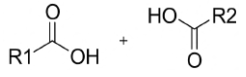
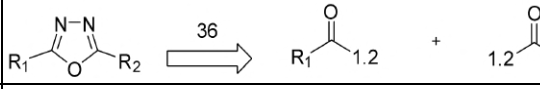
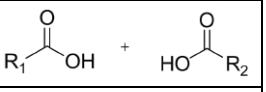
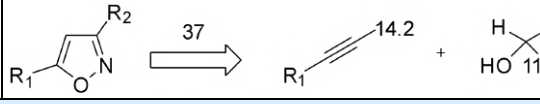
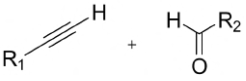
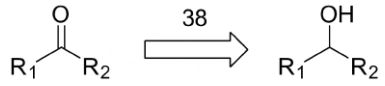
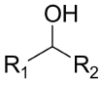
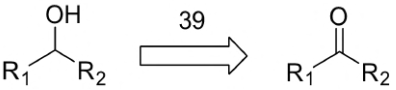
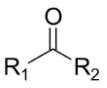
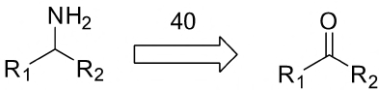
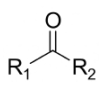
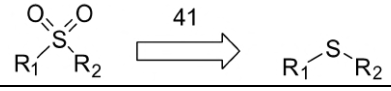
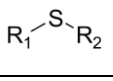
Продолжение на следующей странице

Продолжение таблицы 3

ID	Трансформ	Схема ретро-трансформации	Пример реактантов
R14	Реакция Соногаширы		
R15	Восстановительное аминирование		
R16	Реакция Виттига		
R18.1	S_NAr аминирование o-F-субстратов		
R18.2	S_NAr аминирование п-F-субстратов		
R19	Синтез гидразонов		
R20	Сложноэфирная конденсация		
Ретро-реакции, требующие разрыва более одной связи			
R21	Синтез мочевин		
R22	Синтез карбаматов		
R23	Эпоксидирование по Кори-Чайковскому с раскрытием эпоксидного цикла		
R24	Сопряженное присоединение аминов к α,β-ненасыщенным карбонильным и нитро- соединениям		
R25	Реакция Уги		
R26	Синтез циклических амидинов		
R27	Фосгенирование и сульфо-фосгенирование диаминов		

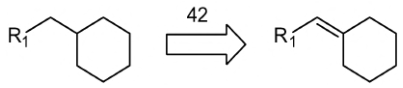
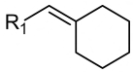
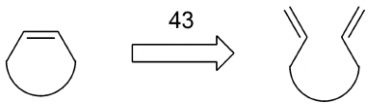
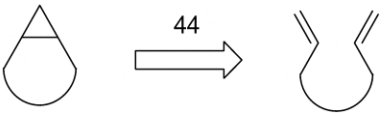
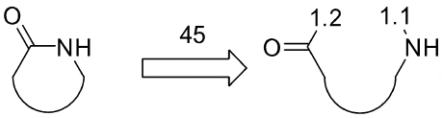
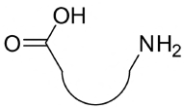
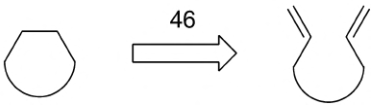
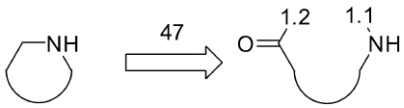
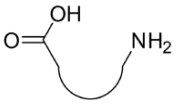
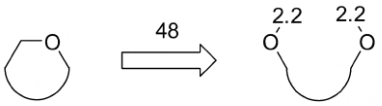
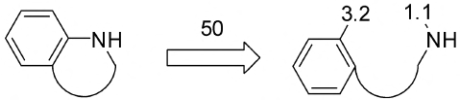
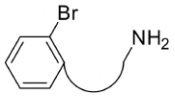
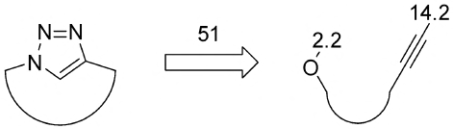
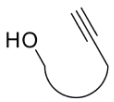
Продолжение на следующей странице

Продолжение таблицы 3

ID	Трансформ	Схема ретро-трансформации	Пример реактантов
R28	Фосгенирование аминоспиртов		
R29	Синтез бензимидазолов и их аналогов		
R30	Синтез бензотиазолов и их аналогов		
R31	Синтез бензоксасолов и их аналогов		
R32	Синтез аминотиазолов		
R33	Синтез тиазолов		
R34	Синтез азидов из первичных спиртов и последующей их циклизацией в 1,2,3-триазолы		
R35	Синтез 1,2,4-оксадиазолов		
R36	Синтез 1,3,4-оксадиазолов		
R37	Синтез изооксазолов		
Квази-однокомпонентные трансформации			
R38	Окисление спиртов в карбонильные соединения		
R39	Восстановление карбонильных соединений до спиртов		
R40	Восстановительное аминирование карбонильных соединений аммиаком		
R41	Окисление сульфидов до сульфонов		

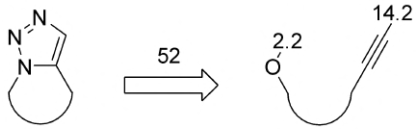
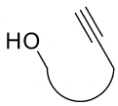
Продолжение на следующей странице

Продолжение таблицы 3

ID	Трансформ	Схема ретро-трансформации	Пример реактантов
R42	Каталитическое гидрирование двойной связи C=C		
Ретро-реакции макроциклизации			
R43	Цикл-замыкающий метатезис		
R44	Цикл-замыкающий метатезис с последующим циклопропанированием двойной связи		
R45	Внутримолекулярный амидный синтез		
R46	Цикл-замыкающий метатезис с последующим восстановлением двойной связи		
R47	Внутримолекулярный амидный синтез с последующим восстановлением амидной связи		
R48	Внутримолекулярная дегидратация диолов		
R49	Внутримолекулярный синтез аминов по Мицунобу из аминоспиртов		
R50	Внутримолекулярное арилирование аминогруппы		
R51	Синтез азидов из первичных спиртов и последующей их внутримолекулярным азид-алкиновым циклоприсоединением в 1,4- замещенные 1,2,3-триазолы		

Продолжение на следующей странице

Продолжение таблицы 3

ID	Трансформ	Схема ретро-трансформации	Пример реактантов
R52	Синтез азидов из первичных спиртов и последующей их внутримолекулярным азид-алкиновым циклоприсоединением в 1,5- замещенные 1,2,3-триазолы		

Важным замечанием к реакциям, представленным в таблице 3 является то, что некоторые из них подразумевают не одну реакцию, а несколько последовательных превращений. Это, например, относится к реакции R34, которая подразумевает, что некоторый первичный спирт будет превращен в азид, например, по реакции Мицунобу, а затем полученный азид вступит в реакцию.

2.2.5 Референсный датасет синтетически релевантных структур

За основу референсного датасета синтетически релевантных структур были взяты следующие датасеты:

1) Датасет биологически-активных молекул ChEMBL v.29 (2 084 724 молекулярных структур) [116];
2) Датасеты от поставщика синтетических малых молекул Enamine: HTS, Advanced, Premium коллекции (2 225 631 молекулярных структур) [117];
3) Датасет зарегистрированных лекарственных веществ и клинических кандидатов (7 378 молекулярных структур).

После объединения датасетов, проведения стандартизации зарядов, удаления дубликатов с точности до стереоизомерии, изотопов, металлсодержащих молекул итоговый датасет содержал 3 319 669 молекулярных структур.

2.2.6 Квази-ретросинтетическая фрагментация и статистический анализ фрагментов

Алгоритм квази-ретросинтетической фрагментации в методе ReRSA основан на алгоритме BRICSDecompose с существенной модификацией, разделяющей отдельные ветви квази-ретросинтетического древа. Каждая ветвь ретросинтетического древа после выполнения всех возможных виртуальных реакций образует набор синтоноподобных фрагментов, каждый из которых уже не может быть подвергнут дальнейшей фрагментации. Такой конечный набор синтоноподобных фрагментов, соответствующий каждой независимой ветви квази-

ретросинтетического древа мы называем **сплитом** (*split* — англ., разбиение). Количество возможных сплитов для молекулярной структуры ограничено сверху. Предполагается, что в 50 первых найденных сплитах найдутся такие сплиты, которые будут синтетически релевантны. То есть сплиты, синтоноподобные фрагменты которых будут индексироваться в референсном датасете фрагментов, а соответствующие синтетические эквиваленты будут индексироваться в базе данных КДИС. Пример того, как разный набор реакций может приводить к разным сплитам продемонстрирован на рис. 12.

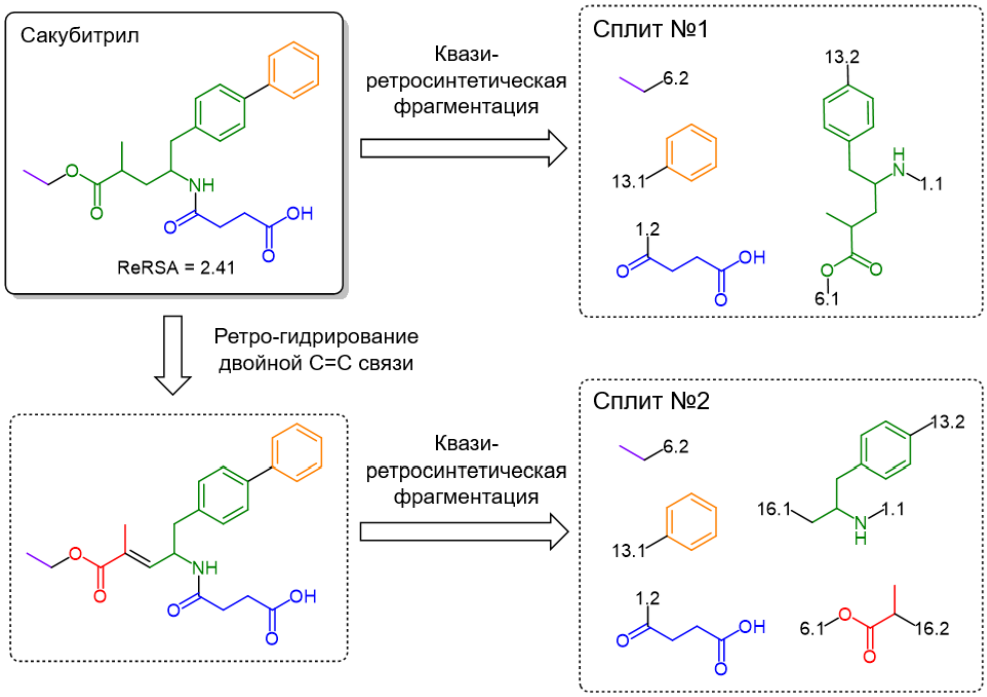


Рисунок 12. Квази-ретросинтетическая фрагментация молекулярной структуры сакубитрила. Продемонстрировано 2 сплита из 18 генерируемых.

В результате квази-ретросинтетической фрагментации получаемые синтоноподобные фрагменты получают метку соответствующего реакционного типа в соответствии с таблицей 3.

Функциональные типы фрагментов отличаются от реакционных типов тем, что они соответствуют функциональным классам органических соединений и не зависят от конкретных ретро-реакций. Так, например, нет принципиальной разницы между молекулой анилина, которая вступает в реакцию ацилирования аминогруппы и молекулой анилина, которая вступает в реакцию алкилирования аминогруппы. Соответственно нет никакой разницы химической разницы между соответствующим молекуле анилина фрагментам 1.1 и 3.1 (см. рис. 13).

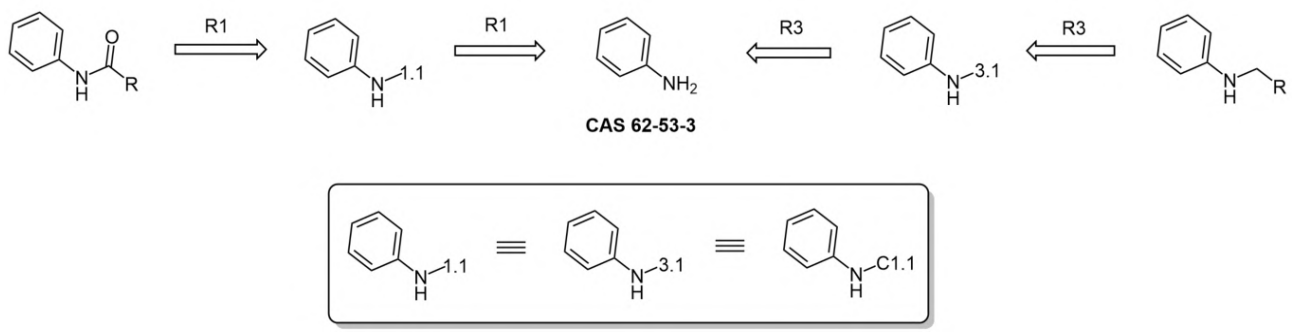
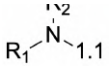
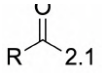
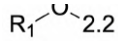
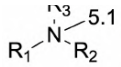
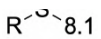
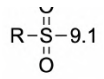
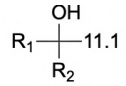
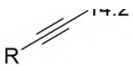
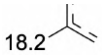
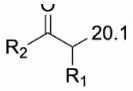


Рисунок 13. Эквивалентность реакционных типов синтоноподобных фрагментов в контексте функциональных типов на примере ретро-ацилирования (R1) и ретро-алкилирования (R3) аминов.

По этой причине для корректного учета синтоноподобных фрагментов в ходе статистического анализа, несколько реакционных типов фрагментов могут быть сгруппированы в один химический тип фрагментов. Группировка реакционных типов в химические типы описана в таблице 4.

Таблица 4. Принцип объединения реакционных типов синтоноподобных фрагментов в функциональные типы

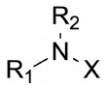
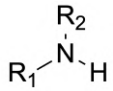
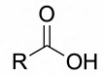
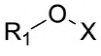
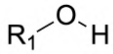
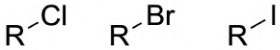
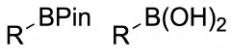
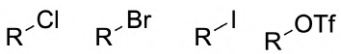
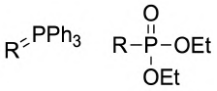
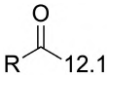
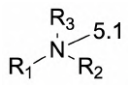
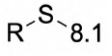
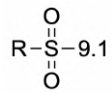
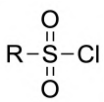
ID	Объединяемые реакционные типы фрагментов	Структура	Функциональный тип
F1.1	1.1, 3.1, 4.3, 10.2, 18.1, 18.3, 19.1		N-нуклеофилы
F1.2	1.2, 2.1, 12.1, 20.2		Ацилирующие агенты
F2.2	2.2, 6.1, 7.3, 13.9		O-нуклеофилы
F3.2	3.2, 4.4, 5.2, 6.2, 7.4, 8.2, 9.2, 11.2, 12.2, 13.1, 13.2, 13.6, 13.7, 13.8, 14.1, 16.2		Галиды и их производные (бороновые кислоты и эфиры, трифлаты, илиды фосфора)
F5.1	5.1		Третичные амины
F8.1	8.1		SH-сульфиды
F9.1	9.1, 10.1		Сульфонил хлориды
F11.1	11.1, 15.2, 16.1, 19.2		Карбонильные соединения
F14.2	14.2		Ацетилены
F18.2	18.2, 18.4		Арил фториды
F20.1	20.1		Метиленовые компоненты

2.2.7 Конвертация синтоноподобных фрагментов в стартовые материалы

Наличие коммерчески доступных исходных соединений (КДИС) в синтетической (соответственно и в ретросинтетической схеме) — важнейшее условие для того, чтобы считать синтетическую доступность для молекулярной структуры более высокой. Более того, чем больше таких стартовых соединений может быть найдено в ходе ретросинтетического анализа, тем выше вероятность синтезировать молекулу. Напротив, если некоторая часть молекулярной структуры не может посредством ретросинтетического анализа найти отображение в пространстве КДИС, то синтетическая доступность по этой причине будет, очевидно, ниже.

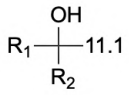
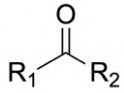
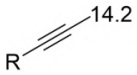
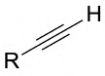
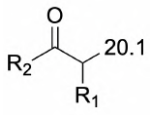
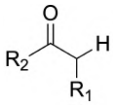
Если обозначить правило, по которому синтоноподобный фрагмент будет превращен в его синтетический эквивалент, то путем сверки синтетического эквивалента с базой КДИС можно учесть наличие или отсутствие синтетического эквивалента в оценке синтезируемости молекулы. В методе ReRSA синтоноподобные фрагменты конвертируются в синтетические эквиваленты (реактанты) согласно правилам, описанным в таблице 5.

Таблица 5. Правила конвертации синтоноподобных фрагментов в синтетические эквиваленты

ID конвертируемых реакционных типов	**Структура**	**Синтетический эквивалент**	**Структура синтетического эквивалента**
1.1, 3.1, 4.3, 10.2, 18.1, 18.3, 19.1		N-нуклеофилы	
1.2, 2.1, 12.1, 20.2		Карбоксильные кислоты	
2.2, 6.1, 7.3, 13.9		Спирты, фенолы	
3.2, 4.4, 5.2, 6.2, 7.4, 8.2, 9.2, 11.2, 12.2		Галиды	
13.1, 13.6, 13.7, 13.8,		Бороновые кислоты, эфиры	
13.2		Галиды, трифоаты	
16.2		Илиды фосфора и фосфонаты	
12.1		Ацилхлориды	
5.1		Третичные амины	
8.1		SH-сульфиды	
9.1		Сульфонил хлориды	

Продолжение на следующей странице

Продолжение таблицы 5

ID конвертируемых реакционных типов	**Структура**	**Синтетический эквивалент**	**Структура синтетического эквивалента**
11.1, 15.2, 16.1, 19.2		Карбонильные соединения	
14.2		Ацетилены	
18.2, 18.4		Арил фториды	
20.1		Метиленовые компоненты	

Конвертация в синтетические эквиваленты частично осуществляется с помощью замены подстроки для экономии времени и упрощения процедуры, частично — с помощью метода RunReactants, где замена подстроки не представляется возможной.

На рис. 14. приведена схема получения конечного набора КДИС для молекулы барицитиниба с использованием правил ретро-трансформаций, описанных в таблице 3, и правил конвертации синтоноподобных фрагментов в КДИС, приведенных в таблице 5.

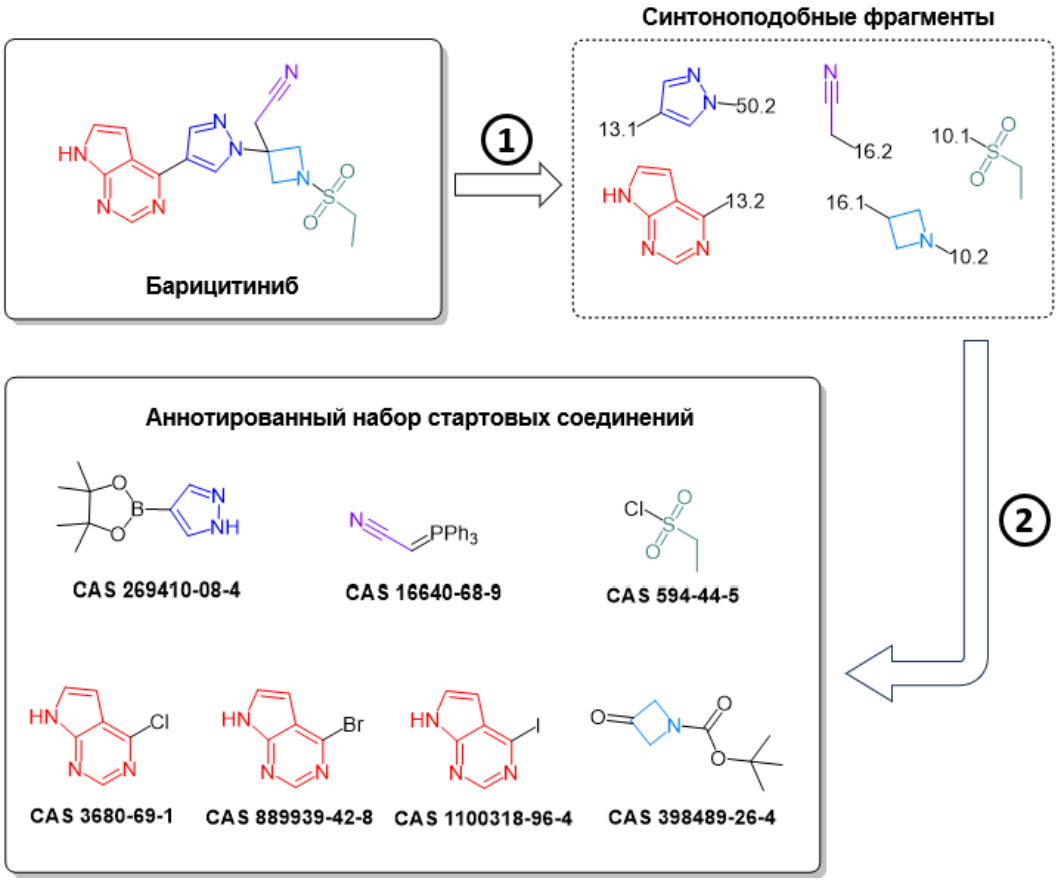


Рисунок 14. Квази-ретросинтетический анализ ингибитора Янус-киназы барицитиниба при помощи алгоритма ReRSA. На этапе ① проходит квази-ретросинтетическая фрагментация молекулярной структуры барицитиниба, в ходе которой получаются синтоноподобные фрагменты. На этапе ② проходит конвертация синтоноподобных фрагментов в синтетические эквиваленты, которые затем индексируются в базе данных коммерчески доступных исходных соединений (КДИС).

2.2.8 Датасет коммерчески доступных исходных соединений

В качестве исходной базы данных коммерчески доступных исходных соединений (КДИС) была взята база данных Enamine Ltd [118]. Общее количество молекулярных структур в исходной базе составило 541 083 единицу. В целях аннотации КДИС независимыми от поставщика идентификационными номерами всем стартовым материалам был присваивался идентификатор системы CAS [119] при помощи ИПП (интерфейс прикладного программирования, англ. — API, *applied programming interface*) на базе ресурса PubChem [108]. Финальное число молекул, для которых удалось найти CAS номера: 478 446.

Особое значение для метода ReRSA представляет преобразование исходной базы данных КДИС в ту, которая должным образом синхронизирована с методом ReRSA. Синхронизация включает в себя, например, преобразование КДИС, содержащих защитные

группы, в те, что защитных групп не имеют. В противном случае, КДИС с защитными группами невозможно было бы учитывать в ходе работы алгоритма ReRSA, поскольку почти все защитные группы могут быть фрагментированы при помощи ретро-реакций метода (табл. 3). Например, Boc-защитная группа фрагментируется при помощи ретро-реакции R22, а бензильная защита аминогруппы фрагментируется при помощи ретро-реакции R3. При преобразовании исходной базы данных КДИС информация о том, что защитная группа была в КДИС, не теряется и хранится в базе данных вместе с продуктом удаления защитной группы, причём поиск синтетических эквивалентов среди стартовых соединений будет проходить по базе депротектированных стартовых соединений со ссылкой на исходное соединение с защитной группой. Например, если исходить из случая квази-ретросинтетического анализа барицитиниба, представленного выше на рис. 14, то процесс конвертации в синтетические эквиваленты с учетом защитных групп можно проиллюстрировать на примере синтоноподобного фрагмента, содержащего азетидиновый цикл (см рис. 15).

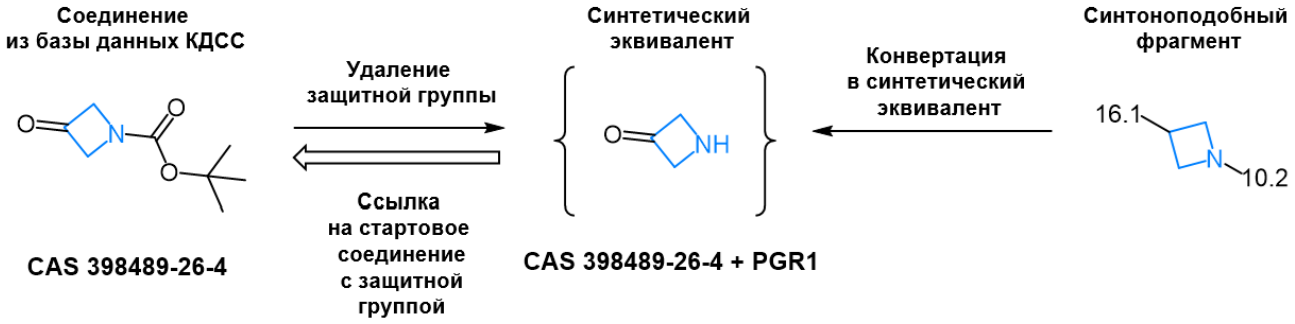
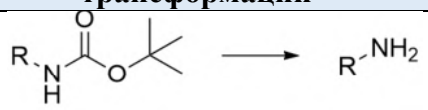
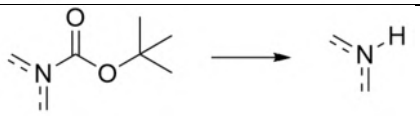


Рисунок 15. Схема процесса снятия защитной группы с КДИС для превращения в синтетический эквивалент, эксплуатируемый алгоритмом ReRSA.

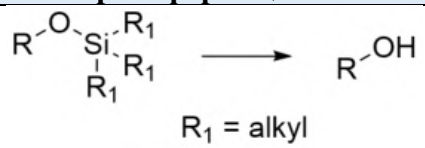
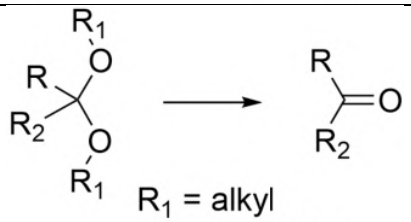
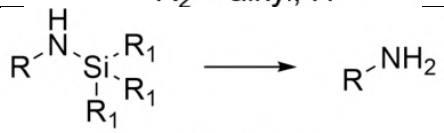
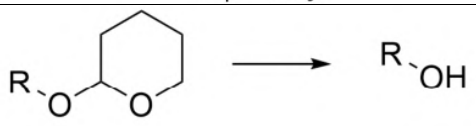
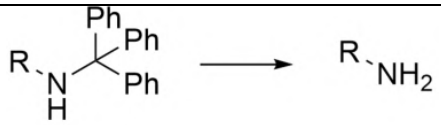
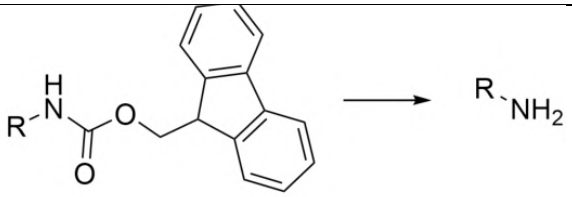
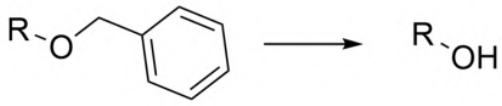
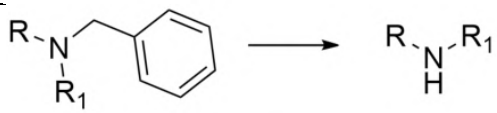
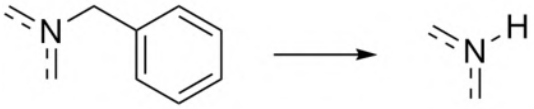
Реакции удаления защитных групп соединений из исходной базы КДИС приведены в таблице 6.

Таблица 6. Реакции удаления защитных групп для преобразования исходной базы данных КДИС для эксплуатации алгоритмом ReRSA

ID	Описание трансформации	Схема трансформации
PGR1	Удаление Boc-группы (защищенный амин)	
PGR2	Удаление Boc-группы (защищенный амин пиррольного типа)	

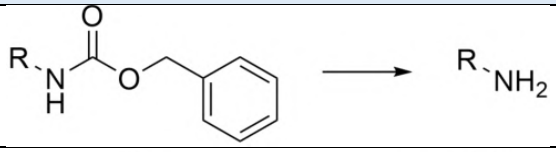
Продолжение на следующей странице

Продолжение таблицы 6

ID	Описание трансформации	Схема трансформации
PGR3	Гидролиз силильных эфиров (защищенный спирт)	 R_1 = alkyl
PGR4	Гидролиз трет-бутилового эфира (защищенная кислота)	
PGR5	Гидролиз ацеталей	 R_1 = alkyl R_2 = alkyl, H
PGR6	Гидролиз аза-силиловых эфиров (защищенный амин)	 R_1 = alkyl
PGR7	Удаление тетрагидропирановой защитной группы	
PGR8	Удаление тритильной защиты	
PGR9	Удаление Fmoc-группы	
PGR10	Удаление бензильной защиты (защищенный спирт)	
PGR11	Удаление бензильной защиты (защищенный амин)	 R_1 = alkyl, H
PGR12	Удаление бензильной защиты (защищенный амин пиррольного типа)	

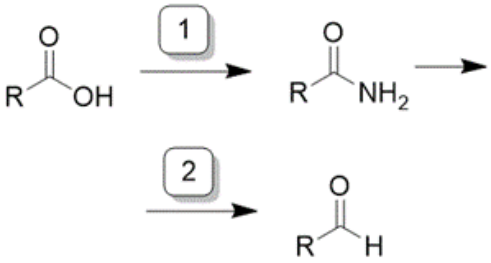
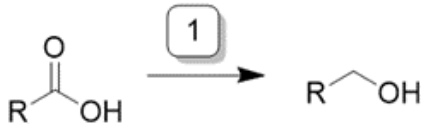
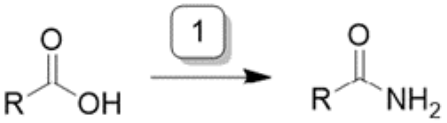
Продолжение на следующей странице

Продолжение таблицы 6

ID	Описание трансформации	Схема трансформации
PGR13	Удаление Cbz-защиты	
PGR14	Деметилирование метоксигруппы для арильных субстратов	

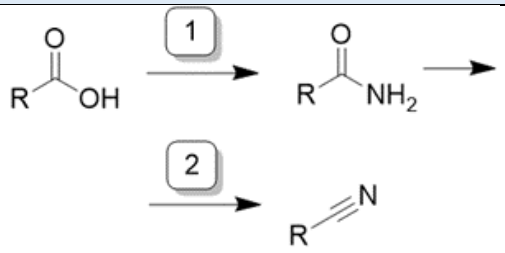
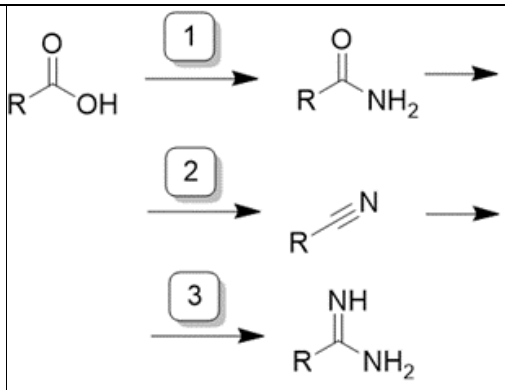
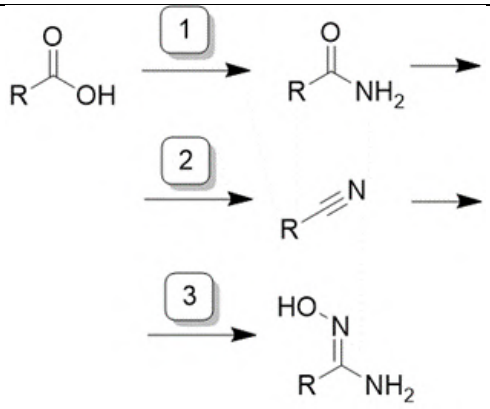
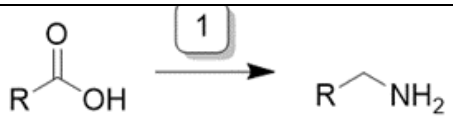
Так же для насыщения базы синтетических эквивалентов исходная база КДИС была расширена при помощи детально описанных реакций, преимущественно преобразующих функциональные группы (FGI, *functional group interconversion*). Реакции, используемые для расширения базы синтетических эквивалентов, описаны в таблице 7.

Таблица 7. Реакции виртуальной конвертации, используемые для подготовки базы КДИС в алгоритме ReRSA

ID	Описание трансформации	Схема трансформации	1) Входной класс 2) Результат конвертации
FGI1	Образование альдегидов из карбоновых кислот через восстановление сложных эфиров		1) Карбоновая кислота 2) Альдегид
FGI2	Восстановление карбоновых кислот до первичных спиртов		1) Карбоновая кислота 2) Первичный спирт
FGI3	Получение первичных амидов из карбоновых кислот		1) Карбоновая кислота 2) Первичный амид

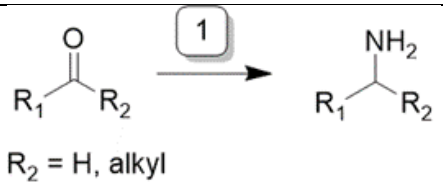
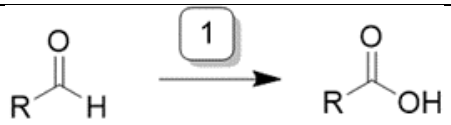
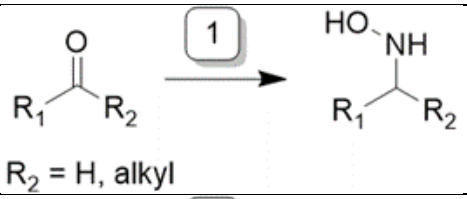
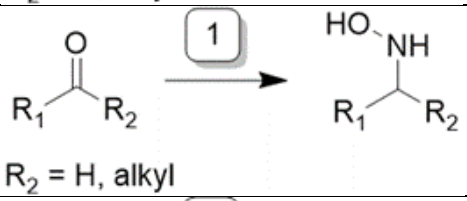
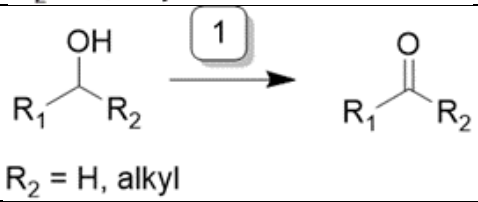
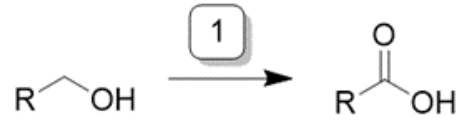
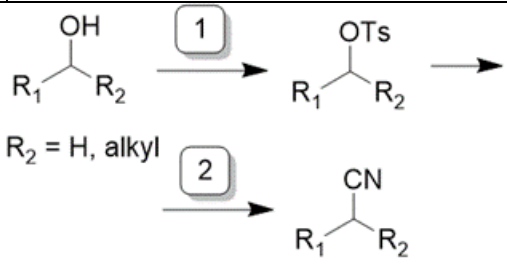
Продолжение на следующей странице

Продолжение таблицы 7

ID	Описание трансформации	Схема трансформации	1) Входной класс 2) Результат конвертации
FGI4	Получение нитрилов из карбоновых кислот через стадию получения первичных амидов		1) Карбоновая кислота 2) Нитрил
FGI5	Получение амидинов из карбоновых кислот через последовательное получение соответствующих амидов и нитрилов		1) Карбоновая кислота 2) Амидин
FGI6	Получение гидроксиамидинов из карбоновых кислот через последовательное получение соответствующих амидов и нитрилов		1) Карбоновая кислота 2)Гидрокси-амидин
FGI7	Синтез ацилхлоридов из карбоновых кислот		1) Карбоновая кислота 2) Ацилхлорид
FGI8	Превращение карбоновых кислот в первичные амины через образование азидов и перегруппировку Курциуса с последующим гидролизом изоцианата		1) Карбоновая кислота 2) Амин

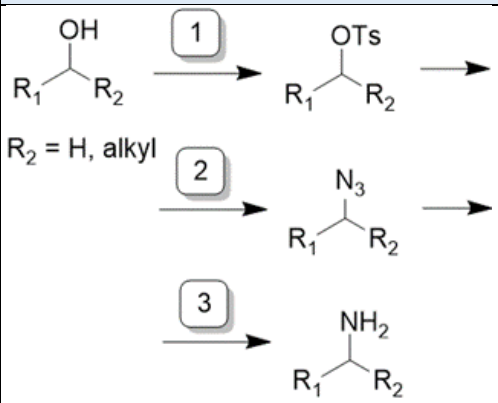
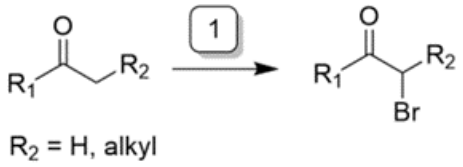
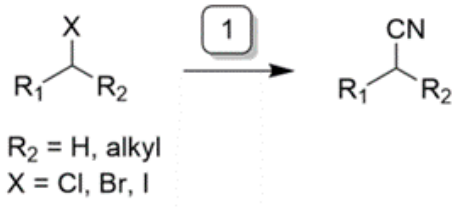
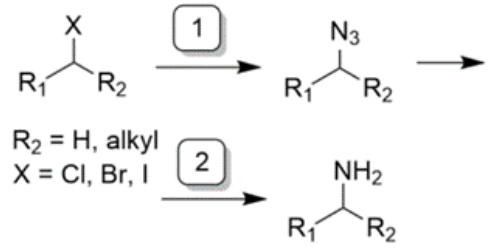
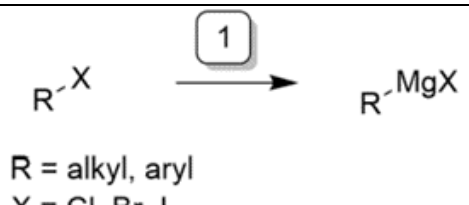
Продолжение на следующей странице

Продолжение таблицы 7

ID	Описание трансформации	Схема трансформации	1) Входной класс 2) Результат конвертации
FGI9	Восстановление альдегидов и кетонов до первичных и вторичных спиртов		1) Карбонильное соединение 2) Спирт
FGI10	Восстановительное аминирование карбонильных соединений		1) Карбонильное соединение 2) Амин
FGI11	Окисление альдегидов до карбоновых кислот по Толленсу		1) Альдегид 2) Карбоновая кислота
FGI12	Синтез оксимов из карбонильных соединений		1) Карбонильное соединение 2) Оксим
FGI13	Синтез алкилбромидов из первичных и вторичных спиртов		1) Спирт 2) Алкилбромид
FGI14	Мягкое окисление первичных и вторичных спиртов до карбонильных соединений		1) Спирт 2) Карбонильное соединение
FGI15	Жесткое окисление первичных спиртов до карбоновых кислот		1) Первичный спирт 2) Карбоновая кислота
FGI16	Тозилирование спиртов с последующим замещением на цианогруппу		1) Спирт 2) Нитрил

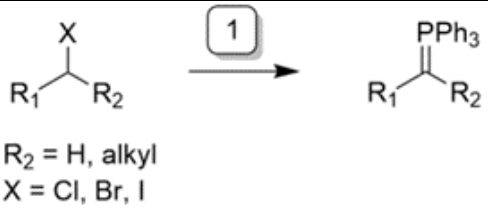
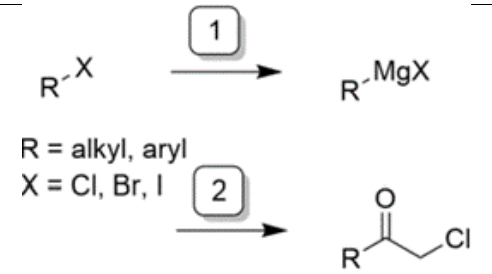
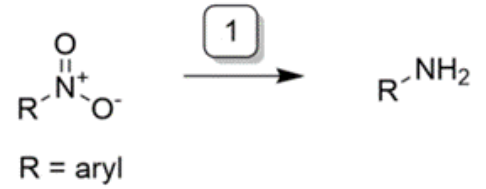
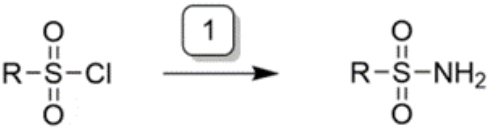
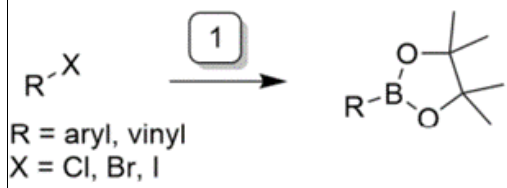
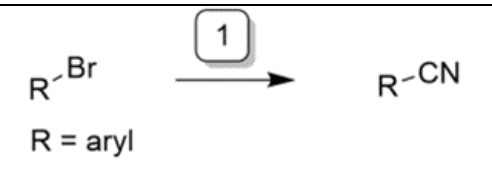
Продолжение на следующей странице

Продолжение таблицы 7

ID	Описание трансформации	Схема трансформации	1) Входной класс 2) Результат конвертации
FGI17	Тозилирование спиртов с последующим образованием азидов и их восстановлением до аминов		1) Спирт 2) Амин
FGI18	Синтез α-бромкетонов		1) Кетон 2) Алкилбромид
FGI19	Нуклеофильное замещение первичных и вторичных алкилгалогенидов на цианогруппу		1) Алкил-галогенид 2) Нитрил
FGI20	Синтез тиолов из акилгалогенидов		1) Алкил-галогенид 2) Тиол
FGI21	Нуклеофильное замещение первичных и вторичных алкилгалогенидов азид-ионом с последующим восстановлением до аминогруппы		1) Алкил-галогенид 2) Амин
FGI22	Получение реактивов Гриньяра из алкил- и арилгалогенидов		1) Алкил- или арилгалогенид 2) Реактив Гриньяра

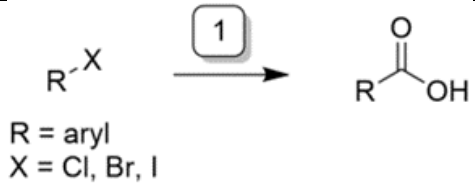
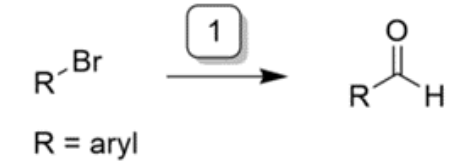
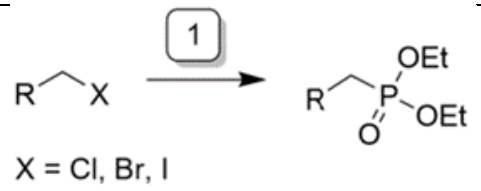
Продолжение на следующей странице

Продолжение таблицы 7

ID	Описание трансформации	Схема трансформации	1) Входной класс 2) Результат конвертации
FGI23	Образование илидов фосфора из алкилгалогенидов		1) Алкил-галогенид 2) Илид фосфора
FGI24	Синтез α-хлор замещенных кетонов из алкилгалогендов через стадию образования реактива Гриньяра и присоединение амида Вайнреба		1) Алкил-галогенид 2) α-хлоркетон
FGI25	Восстановление нитрогруппы по Зинину		1) Нитро-соединение 2) Амин
FGI26	Синтез сульфонамидов из сульфонилхлоридов		1) Сульфонил-хлорид 2) Сульфонамид
FGI27	Синтез эфиров пинаколборана из арил- и винилгалогенидов по Мияуре		1) Галогенид 2) Пинаколовый эфир
FGI28	Каталитическое аминирование арилхлоридов и арилиодидов по Бухвальду-Хартвигу		1) Арилгалогенид 2) Амин
FGI29	Каталитическое цианирование арилбромидов по Розенмунду-фон Брауну		1) Арилбромид 2) Нитрил

Продолжение на следующей странице

Продолжение таблицы 7

ID	Описание трансформации	Схема трансформации	1) Входной класс 2) Результат конвертации
FGI30	Каталитическое карбоксилирование арилгалогенидов		1) Алкил-галогенид 2) Карбоновая кислота
FGI31	Каталитическое формилирование арилбромидов		1) Арилбромид 2) Альдегид
FGI32	Синтез реактива Гриньяра из алкинов		1) Алкин 2) Реактив Гриньяра
FGI33	Получение диэтилфосфонатов по Михаэлису-Арбузову		1) Алкил-галогенид 2) Диэтилфосфонат

При этом стоит опять же отметить, что в некоторых случаях подразумевается последовательность нескольких химических реакций, которая приводит к трансформации. Необходимость включения некоторого метода FGI в таблицу определялась двумя факторами:

1. В исходной базе КДСМ было относительно небольшое количество представителей некоторых функциональных классов химических соединений;
2. Анализ синтезов зарегистрированных лекарственных веществ продиктовывал насыщение библиотеки синтетических эквивалентов определенными функциональными классами.

Например, α-хлоркетоны являются важными прекурсорами в синтезе гетероциклических соединений, таких как тиазолы и оксазолы. Тем не менее, в исходной базе КДИС присутствует всего 645 α-хлоркетонов. Распространенным способом синтеза α-хлоркетонов является реакция арилмагнийгалогенидов с амидом Вайнреба α-хлоруксусной кислоты (FGI24). Благодаря дополнительной генерации α-хлоркетонов удалось получить дополнительно почти 14 тысяч молекулярных структур синтетических эквивалентов со ссылками на исходные КДИС. Этот и другие примеры дополнительной генерации синтетических эквивалентов приведены в таблице 8.

Таблица 8. Пример генерации синтетических эквивалентов для 4 функциональных классов (число сгенерированных структур приводится без учета последующей процедуры удаления дубликатов для финального подготовленного датасета)

Класс	**Количество в исходном датасете**	**Общее число молекул, полученное в ходе постконвертации**	**Индексы схем генерации (см. табл. 7)**
Альдегиды	8175	67812	FGI1, FGI14, FGI31
Альфа-бром кетоны	532	3298	FGI18
Альфа-хлор кетоны	645	14134	FGI24
Диэтилфосфонаты	66	8578	FGI34

2.2.9 Фильтрация синтетически нерелевантных подструктур[10]

Помимо синтоноподобных фрагментов, получаемых в ходе квази-ретросинтетической фрагментации, особое внимание было уделено 5-членным ароматическим гетероциклам, представляющих проблему при анализе синтезируемости молекулярных структур. При фрагментации стратегическая связь между линкером и циклом или двумя циклами может быть разорвана, что подразумевает, например, трансформ реакции Гриньяра, или кросс-сочетания Сузуки, или Бухвальда-Хартвига. Такой подход позволяет, как было описано выше, получить предполагаемые стартовые материалы, однако никак не учитывает синтетическую осуществимость цикла с конкретными заместителями. Например, богатые электронами ароматические кольца оксазолов с донорными заместителями реакционноспособны и склонны к гидролизу, что, очевидно, понижает их синтетическую доступность. Для устранения этого недостатка был предложен алгоритм автоматической фильтрации молекулярных структур, содержащих синтетически нерелевантные подструктуры, с использованием иерархически организованной библиотеки SMARTS-подструктур, описывающих пятичленные ароматические гетероциклы, включая первый атом заместителей. Синтетически нерелевантной подструктурой в данном случае следует называть те подструктуры, которые не могут быть обнаружены в химических соединениях, представленных в референсных базах

[10] При работе над данным разделом диссертации использованы материалы публикации Бондарев Н., **Загрибельный Б.**, Федорченко С.А., Иваненков Я.А., Палюлин В.А. Моделирование синтетической доступности потенциальных лекарственных веществ, содержащих пятичленные ароматические гетероциклы // Известия Академии наук. Серия химическая. — 2025. — Т. 74. — № 6. — С. 1687–1703. (Переводная версия: Bondarev N., **Zagribelyy B.**, Fedorchenko S.A., Ivanenkov Ya. A., Palyulin V.A. Modeling of synthetic accessibility of potential drug molecules containing five-membered aromatic heterocycles // Russian Chemical Bulletin. — 2025. — Т. 74. — № 6. — С. 1687–1703.)

данных, а значит, молекулярные структуры, содержащие эти фрагменты должны быть отсеяны, исходя из бизнес-логики генеративного эксперимента.

В качестве основы фильтрующего алгоритма предлагаются иерархически организованные библиотеки подструктур, учитывающие локальное окружение фрагмента, в данном случае — пятичленных гетероциклов. Иерархическая организация библиотеки позволяет значительно сократить время поиска конкретной подструктуры в молекулярном графе.

Для описания фрагментов был выбран язык под-структур SMARTS (SMILES arbitrary target specification) [120], который является наиболее гибким и одновременно достаточно простым в использовании строковым представлением молекулярных структур и фрагментов. Язык SMARTS допускает использование логических операторов И, ИЛИ. Например, для описания трех бензилгалогенидов (иодида, бромида и хлорида) понадобится три SMILES-строки [102], в то время как SMARTS позволяет записать надструктуру, объединяющую 3 подструктуры, одной строкой (см. рис. 15). Возможность существенного сокращения числа записей является большим преимуществом при создании больших библиотек.

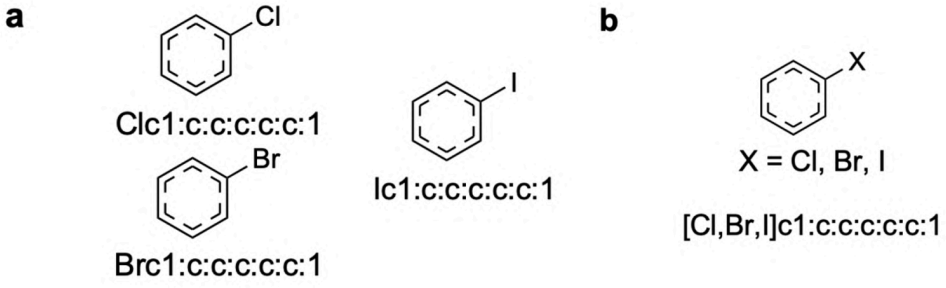


Рисунок 15. Сравнение представлений молекулярных структур SMILES (**a**) и SMARTS (**b**).

2.2.9.1 Генерация иерархической библиотеки фрагментов

Генерацию иерархической библиотеки возможных с точки зрения валентности SMARTS-подструктур проводили заменой символов в SMARTS-строке. Такой подход позволяет почти не прибегать к операциям с молекулярными объектами. Генерация производилась на языке программирования Python с использованием встроенных модулей pandas, numpy, itertools, а также хемоинформатического пакета RDKit с открытым исходным кодом [27].

Первым этапом стало определение нулевого слоя — базового ароматического пятичленного цикла (рис. 16 а), и задание азбуки элементов. Библиотека была ограничена пятичленными гетероциклами, содержащими до трех гетероатомов. Таким образом, азбука состоит из 5 элементов (табл. 9). Также важно отметить, что азбука элементов учитывает

“свободную валентность” атома, входящего в состав гетероцикла, если таковая имеется, в виде символа “*”. Это необходимо для того, чтобы на место символа “звездочки” в дальнейшем вставить атом заместителя или водород.

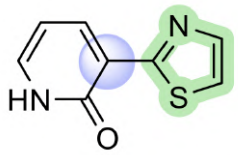
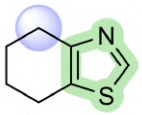
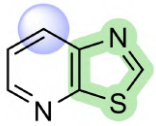
Таблица 9. Азбука элементов гетероциклического ядра

Атом	**SMARTS**	**Визуализация циклического фрагмента, включающего описываемый атом**
Углерод	[#6](*)	
Замещенный азот	[#7+0](*)	
Азот	[#7+0]	
Кислород	[#8]	
Сера	[#16]	

Комбинаторная замена символа “a” в строке нулевого слоя на атомные примитивы из азбуки элементов, позволила получить первый слой библиотеки, состоящий из 7 элементов, каждый из которых соответствует пятичленному ароматическому гетероциклу с конкретным числом и положением гетероатомов (рис. 16.b). Таким же образом, заменяя “[a:1]” на атомные примитивы, получили 3-й слой библиотеки, содержащий 21 элемент — конкретный гетероцикл (рис. 16.c). При этом ввиду того, что одну подструктуру можно записать несколькими разными строками на SMARTS, возникали дублирующие друг друга подструктуры. Чтобы идентифицировать и удалить дубликаты, мы конвертировали SMARTS-строку в молекулярный объект RDKit и выполняли поиск изоморфного подграфа (стандартный метод Has-SubstructMatch библиотеки RDKit), при полном совпадении молекулярных графов, одну из строк удаляли.

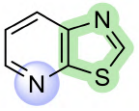
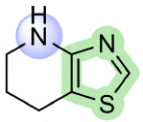
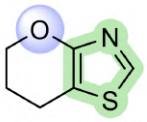
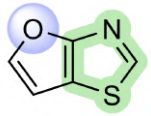
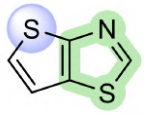
Следующий этап — задание азбуки заместителей. Мы определили 18 типов релевантных с точки зрения медицинской химии заместителей, каждый из которых представляет собой один специфицированный атом (табл. 10). Заместитель характеризуется примитивом связи и атомным примитивом. Это необходимо, чтобы отличать, например C_3 и C_5 — ароматический атом углерода и ароматический атом углерода в конденсированном цикле, у которых одинаковые атомные примитивы. Сначала, заменяя символ “*” на символ атома водорода в строках второго слоя, и учитывая принятое ограничение на наличие в подструктуре от 2 до 3 заместителей, мы сформировали третий слой. Этот слой состоит из 86 элементов и представляет собой набор гетероциклов второго слоя с двумя или тремя заместителями с учетом их положений (рис. 16.d).

Таблица 10. Типы заместителей, использованные при создании иерархической библиотеки SMARTS-подструктур пятичленных гетероциклов, учитывающих локальное окружение гетероцикла (гетероциклическое ядро, относительно которого рассматривается заместитель, подсвечено зеленым цветом)

Обозначение	**SMARTS-представление связи**	**SMARTS-представление атома**	**Описание**	**Пример функциональной группы**
C_1	!=!@	[CH3,CH2]	Алифатический углерод с двумя или тремя атомами водорода	$-CH_3$, $-CH_2R$
C_2	!=!@	[CH,CH0]	Алифатический углерод с одним атомом водорода или без водородов.	$-CHR_2$, $-CR_3$, -COR, -CN
C_3	!=!:	[c]	Ароматический углерод	
C_4	!=@	[C]	Алифатический углерод в конденсированном цикле	
C_5	:	[c]	Ароматический углерод в конденсированном цикле	
N_1	!=!@	[NH2,NH,N+]	Алифатический азот с двумя или тремя атомами водорода, заряженный алифатический азот	$-NH_2$, -NHR, $-NO_2$, $-N^+R_4$
N_2	!=!@	[N+0H0]	Алифатический азот без атомов водорода	$-NR_2$, -N=R

Продолжение на следующей странице

Продолжение таблицы 10

Обозначение	**SMARTS-представление связи**	**SMARTS-представление атома**	**Описание**	**Пример функциональной группы**
N_3	!=!:	[n+0]	Ароматический азот	
N_4	:	[n+0]	Азот в ароматическом конденсированном цикле	
N_5	!=@	[N+0]	Алифатический азот в конденсированном цикле	
O_1	!=!@	[O]	Кислород	-OH, -OR
O_2	!=@	[O]	Кислород в конденсированном цикле	
O_3	:	[o]	Кислород в ароматическом конденсированном цикле	
S_1	!=!@	[S]	Сера	-SH, -SR, -SO_2R
S_2	!=@	[S]	Сера в конденсированном цикле	
S_3	:	[s]	Сера в ароматическом конденсированном цикле	
F		F	Фтор	-F
Cl		Cl	Хлор	-Cl

После комбинаторного перебора и подстановки всех типов заместителей было получено 52 896 SMARTS-подструктур, которые составили четвертый, конечный слой (рис. 16.е). После удаления дубликатов (симметричных подструктур) осталось 49 152 SMARTS-строки. Таким образом была построена иерархическая библиотека подструктур, где каждый слой представляет набор валидных SMARTS-строк, а каждая ветвь соответствует уточнению

подструктуры: от любого пятичленного ароматического цикла на нулевом слое, до конкретного гетероцикла с конкретными типами заместителей в конкретных положениях на четвертом слое. Фрагмент библиотеки представлен в Приложении А.

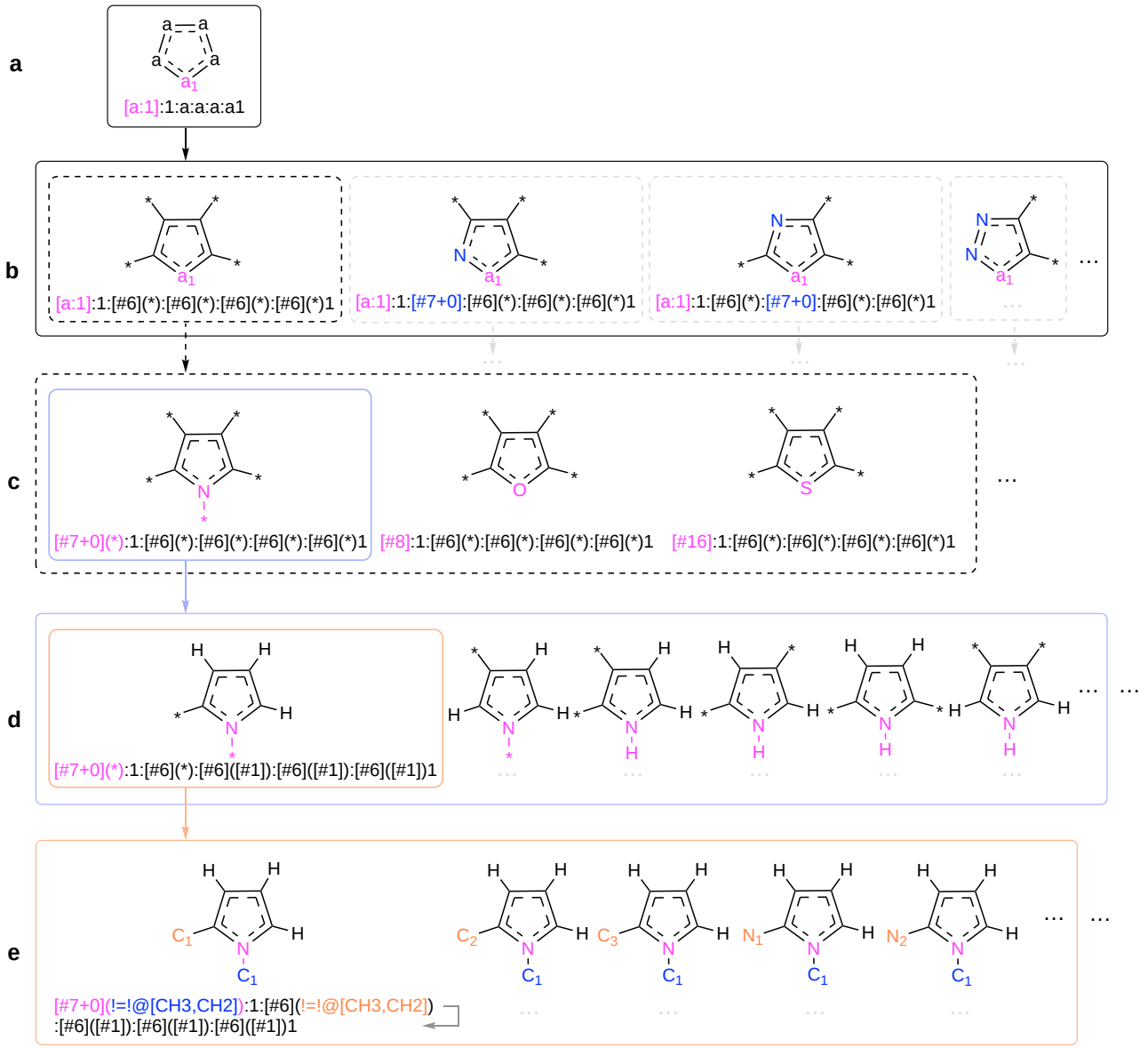


Рисунок 16. Схема генерации иерархической библиотеки подструктур пятичленных ароматических гетероциклов. Здесь и далее перенос SMARTS на новую строку обозначен символом ↲.

2.2.9.2 Сбор фрагментных статистик.

В качестве референсных наборов синтезированных соединений были выбраны открытая база данных ChEMBL 29 [116] — 1 850 431 уникальных структур, и база коммерчески доступных соединений для скрининга Enamine [117] (HTS, Advanced, Premium Collections) — 2 823 257 уникальных соединений после стандартного препроцессинга (стандартизация структур, удаление солевой части, удаление дубликатов). Анализ статистик

встречаемости подструктур проводили независимо по двум выборкам с помощью стандартного поиска изоморфного подграфа (табл. 11).

Таблица 11. Число обнаружений подструктур в обучающих наборах

Число обнаружений	Число подструктур в наборе	
	ChEMBL 29	Enamine
≥ 1	5 023	3 069
≥ 10	2 440	1 891
< 10	46 712	47 261

В качестве базовой гипотезы о природе синтетической доступности было выдвинуто положение о том, что наличие редкой подструктуры в молекулярной структуре снижает вероятность того, что молекула является синтетически осуществимой. Был выбран порог встречаемости подструктур на уровне 10, поскольку подструктуры, встречающиеся реже, могут представлять собой редкие, нетипичные структуры, что может вызвать подозрения относительно их синтетической доступности и реалистичности. Такой порог позволяет исключить потенциально ошибочные или артефактные подструктуры.

2.2.9.3 Оптимизация библиотеки подструктур.

Ввиду большого числа штрафующих фрагментов была реализована автоматическая процедура объединения однородных заместителей с получением консенсусной SMARTS-надструктуры из нескольких SMARTS-строк с помощью регулярных выражений на Python. Данная процедура позволила сократить число штрафующих SMARTS-подструктур с 46 250 до 7 157 штук. Пример объединенной SMARTS-подструктуры представлен на рис. 17. Максимально удавалось объединить 12 подструктур в одну SMARTS-строку. Стоит отметить, что ранее в профильной литературе подобные операции над SMARTS-строками в автоматизированном режиме описаны не были. Следует отметить, что усложнение SMARTS-объекта практически не влияет на время выполнения поиска изоморфного подграфа. Основное время затрачивается на наложение графа, тогда как сравнение атомных примитивов из списка выполняется почти мгновенно.

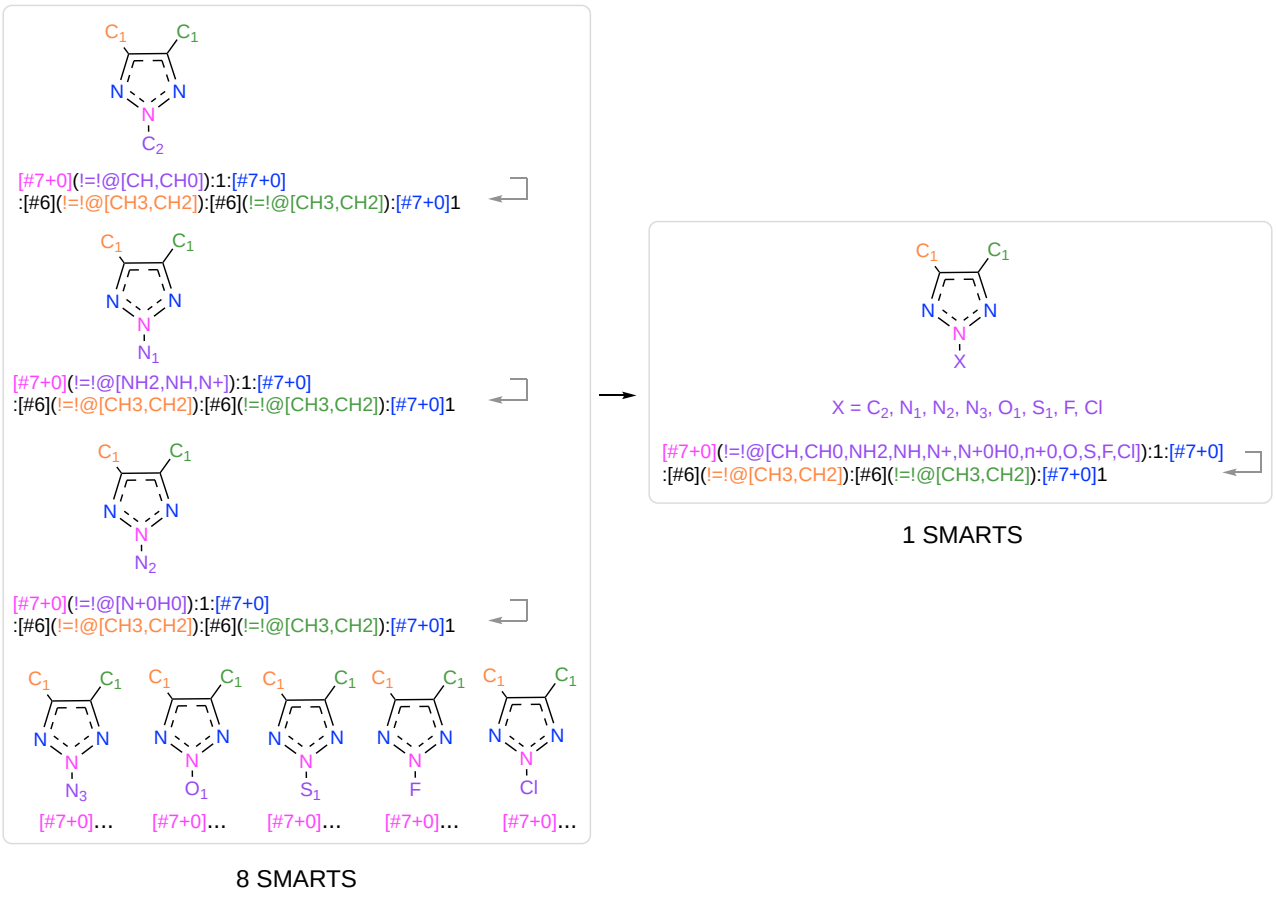


Рисунок 17. Визуализация SMARTS-подструктуры после объединения заместителей. Условные обозначения типов заместителей приведены в табл. 10.

2.2.9.4 Иерархический алгоритм фильтрации.

Иерархическая структура библиотеки штрафующих SMARTS-подструктур позволила разработать иерархический алгоритм их поиска в молекулярных структурах. Алгоритм представляет собой классический поиск по дереву, который завершается при первой найденной подструктуре конечного слоя библиотеки (см. рис. 18). На вход подается SMILES-строка. Вначале проверяется наличие пятичленного ароматического гетероцикла; при его отсутствии алгоритм завершает работу, и молекулярная структура успешно проходит фильтр. Если гетероцикл обнаружен, алгоритм проверяет наличие первой подструктуры из первого слоя библиотеки. При положительном результате поиск продолжается на следующем слое (на четвертом, последнем слое, структура отбраковывается, и алгоритм завершается), при отрицательном — переход к следующей подструктуре того же слоя (смена ветви). Если все подструктуры слоя не были обнаружены (например, пятичленный ароматический гетероцикл не соответствует подструктурам из библиотеки), алгоритм завершает работу без наложения штрафа.

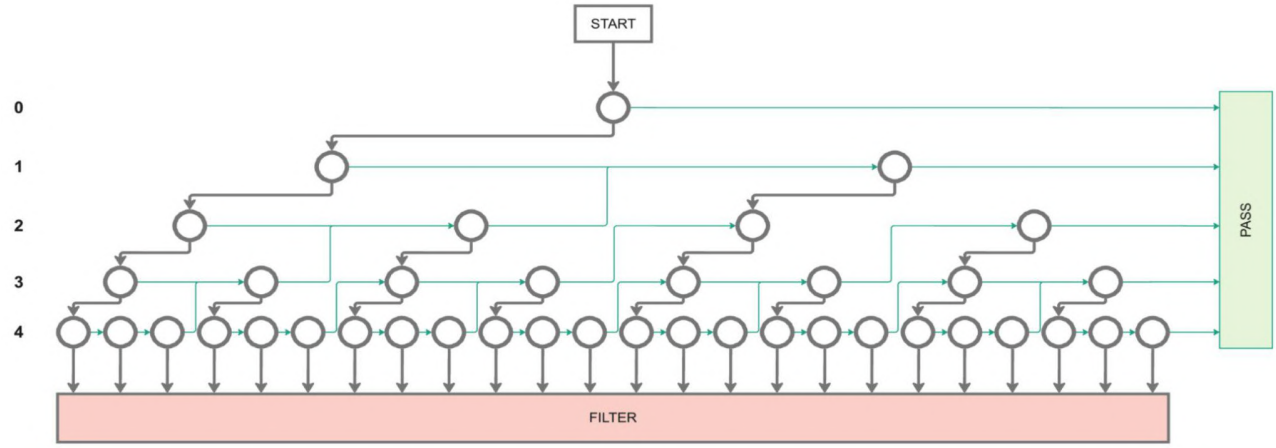


Рисунок 18. Схема иерархического алгоритма поиска подструктуры. Каждый узел соответствует выполнению процедуры поиска изоморфного подграфа для соответствующей подструктуры из библиотеки. В случае успеха (подструктура найдена) осуществляется переход на следующий слой (серая стрелка). В противном случае осуществляется переход на соседний узел или следующий узел предыдущего слоя (зеленая стрелка). Попадание в красный прямоугольник “FILTER” означает, что молекулярная структура не прошла фильтрацию и содержит нерелевантный фрагмент, попадание в зеленый прямоугольник “PASS” означает, что молекулярная структура успешно прошла фильтр.

Такой подход в сочетании с сокращением библиотеки за счет объединения однородных заместителей позволили значительно (более чем в 250 раз) повысить производительность фильтрации молекулярных структур по сравнению с алгоритмом последовательного перебора (см. табл. 12). Указанной производительности более чем достаточно для фильтрации молекулярных структур, производимых современными генеративными моделями.

Таблица 12. Сравнение производительности алгоритмов фильтрации синтетически нерелевантных молекулярных структур для набора из 30 тысяч сгенерированных молекулярных структур

Алгоритм фильтрации	**Последователь-ный перебор**	**Иерархичес-кий поиск**	**Иерархический поиск с кластеризованным набором**
Число SMARTS-подструктур	46 250	46 250	7 157
Время работы алгоритма на 1 CPU, чч:мм:сс	1:22:45	0:00:50	0:00:20
Производительность алгоритма на 1 CPU, структур/с	~6	600	1 500

2.2.10 Функция ReRSA для агрегации факторов, влияющих на синтезируемость

Учитывая всё вышесказанное и принципиальную схему метода ReRSA (см. рис. 9), была предложена следующая логика для агрегации факторов, ассоциированных с синтетической доступностью молекулярных структур, которая будет возвращать значение оценки синтезируемости.

На основании разбиения референсного датасета молекул были собраны частоты (*fr*) встречаемости каждого синтона подобного фрагмента — это отношение числа молекул, в которых присутствует фрагмент, к количеству всех молекул в референсном датасете. Согласно определению, частоты встречаемости лежат в диапазоне от 0 до 1. Чем меньше частота некоторого фрагмента, тем выше должно быть значение ReRSA. Поэтому для удобства и большей интерпретируемости брали обратный логарифм от (*fr*) и определяли *fr′*:

$$fr' = 1 - \lg(fr).$$

Чем выше *fr′*, тем выше должна быть итоговая оценка, т. е. вероятность безуспешного синтеза молекулярной структуры тем выше, чем выше *fr′*. Далее для каждого фрагмента, встреченного в референсном наборе молекул, был рассчитан структурный дескриптор (*sd*), определяющий структурную сложность и рассчитываемый по формуле (2):

$$\begin{aligned} sd = (&ChiralCentersCount + RingCount + RingSideChainsCount \\ &+ spiroCount + BiggestRingSize + FusedRingCount + BridgeAtomsCount \\ &+ \frac{HeavyAtomCount}{MW}\Big) \cdot (Q^1 + MW), \end{aligned} \tag{2}$$

где

- *ChiralCentersCount* — число хиральных атомов углерода,
- *RingCount* — число циклов,
- *RingSideChainsCount* — число боковых цепей, присоединенных к циклам,
- *SpiroCount* — число спиро-атомов углерода,
- *BiggestRingSize* — число атомов в самом большом цикле, если оно больше 6; в противном случае — 0,
- *FusedRingsCount* — количество конденсированных циклов,
- *BridgeAtomsCount* — число атомов во главе моста в бициклических системах,
- *HeavyAtomCount* — число тяжелых атомов,
- *MW* — молекулярный вес,

- Q^1 — нормализованный квадратичный индекс 1, вычисляемый как $Q^1 = 3 - 2 \cdot A + Z1/2$, где A — количество тяжелых атомов, $Z1$ — первый индекс Загреба [121].

Финальный дескриптор фрагмента (*SD*), учитывающий его распространенность в референсном наборе молекул, структурную сложность (2), а также возможность конвертации в КДИС, рассчитывается для всех фрагментов, присутствующих в референсном наборе молекул, по формуле (3а):

$$\mathrm{SD} = \begin{cases} \dfrac{sd \cdot fr'}{2}, & \text{если фрагмент конвертируется в КДИС;} \\ sd \cdot fr', & \text{если фрагмент не конвертируется в КДИС.} \end{cases} \quad (3а)$$

Рассчитанные *SD* для всех фрагментов хранятся в словаре и используются для расчета ReRSA для молекулярной структуры, после разбиения ее на фрагменты. В случае, если был выбрана “мягкая” политика расчета (*SOFT policy*), и при фрагментации молекулярной структуры был идентифицирован фрагмент, которого нет в референсном наборе ($fr = 0$), его *SD* определяется по формуле (3б):

$$\mathrm{SD} = \begin{cases} sd \cdot (1 - \lg(fr_{cp})), & \text{если } fr = 0, \text{ и фрагмент конвертируется в КДИС;} \\ sd \cdot 100, & \text{если фрагмент не конвертируется в КДИС.} \end{cases} \quad (3б)$$

где $fr_{\text{ср}}$ – среднее частот всех фрагментов, присутствующих в референсном наборе.

В случае применения “жесткой” политики расчета (*strict policy*) финальный скор сплита (*FSS*), содержащего фрагмент, отсутствующий в референсном наборе и не конвертируемый в КДИС, принимается равным 10. В обратном же случае *FSS* рассчитывается на основании нижеследующих формул.

Первичный скор сплита (*PSS, primary split score*) является суммой финальных дескрипторов (*SD*) всех фрагментов, входящих в состав сплита, рассчитанных по формулам (3а) или (3б):

$$PSS = \sum_{i}^{N} SD_i.$$

PSS может принимать значения от нуля до бесконечности. Чтобы сделать скор сплита более удобным в использовании, можно использовать множество нормализующих функций.

Например, если желаемое значение счета должно быть между нулем и единицей, то можно использовать сигмоидальную функцию. Чтобы получить счет в определенном диапазоне, можно, например, применить функцию арктангенса с некоторыми параметрами, специфичными для диапазона. В случае арктангенса *PSS* подвергается следующему математическому преобразованию:

$$TSS = \arctan\left(\frac{PSS}{1000}\right) \cdot \frac{2}{\pi} \cdot 9 + 1.$$

Таким образом мы получаем нормализованный скор сплита (*TSS*). Финальный скор сплита (*FSS, final split score*), учитывающий число фрагментов в сплите, а также число успешно конвертированных в КДИС фрагментов, рассчитывается по формуле:

$$FSS = \begin{cases} X \cdot TSS^{\frac{1}{1+\frac{n}{N}+0.5}} \cdot \frac{1}{n}, & \text{если } n = N; \\ X \cdot TSS^{\frac{1}{1+\frac{n}{N}}} \cdot \frac{1}{n}, & \text{если } n < N, \end{cases}$$

где *n* — число успешно конвертированных в КДИС фрагментов, *N* — общее число фрагментов в сплите, *X* — поправка в случае применения трансформа Уги. Соответственно, чем выше *n*, тем ниже *FSS* и ниже ReRSA для молекулярной структуры, а если же *n = N,* то молекулярная структура награждается еще более низким значением ReRSA. Помимо награды за полную конверсию сплита, была предусмотрена награда сплита за применение многокомпонентных реакций в ходе квазиретросинтетической фрагментации. Многокомпонентные реакции — мощный инструмент для быстрого построения структурно-сложных каркасов молекул [122]. В связи с этим вклад возможности использования многокомпонентных реакций следует считать положительным при оценке синтезируемости. Указанная поправка вносит вклад в финальный скор сплита: в случае, если для получения сплита была применена реакция Уги, $X = 0.8$, в остальных случаях $X = 1$.

Финальные скоры всех сплитов агрегируются в финальный ReRSA скор, который затем нормируется в диапазоне от 1 до 10:

$$\mathrm{ReRSA} = \left(M \left(FSS_{\mathrm{MIN}} - \frac{1}{1^K} \sum_i^K \frac{1 - \mathrm{sim}(split_i \text{ to } split_{\mathrm{MIN}})}{1.1^{FSS_i - FSS_{\mathrm{MIN}}}} \right) - 1 \right) \cdot 1.7 + 1,$$

где *K* – число сплитов для молекулярной структуры, FSS_{MIN} — минимальный финальный скор из всех сплитов, $sim(split_i\ to\ split_{MIN})$ — оценка схожести сплита по Танимото по отношению к сплиту с минимальным *FSS, M* — штраф за наличие макроцикла в молекулярной

структуре, $M = 1.5$, в случае наличия макроцикла, и $M = 1$ в обратном случае. Таким образом, основной вклад в итоговое значение вносит сплит в наименьшим *FSS*. Однако, остальные сплиты также вносят понижающий вклад в оценку, который тем больше снижает итоговое значение ReRSA, чем меньше его *FSS* и чем больше его отличие от лучшего сплита. Помимо этого, значение ReRSA для структур, содержащих макроциклы должно быть заведомо выше, поскольку выходы реакций макроциклизации далеко не всегда оптимальны [123]. Это и учитывает поправочный коэффициент *M*.

Важно подчеркнуть, что учет наличия в молекулярной структуре синтетически нерелевантного фрагмента осуществляется модулем, описанном в разделе 2.2.9, только в случае применения “жесткой” политики расчета ReRSA. Проверка на наличие такого фрагмента осуществляется прежде вышеописанных расчетов, и, если такой фрагмент обнаружен, для такой молекулярной структуры возвращается ReRSA = 10, и фрагментация с последующим подсчетом финального скора не запускается. При применения “мягкой” политики расчета, модуль проверки на наличие синтетически нерелевантных фрагментов не запускается, и расчет проходит согласно вышеописанной схеме.

3. Результаты и их обсуждение[11]

3.1 Платформа генеративной химии Chemistry42 как интегрированное решение для автоматизированного моделирования структур потенциальных лекарственных веществ.

3.1.1 Историческое развитие идеи о платформе генеративной химии

Успех подхода, предложенного в статье нашей научной группы, посвященной генеративному дизайну ингибиторов DDR1 киназы при помощи ГО на основе генеративной

---

[11] При работе над данным разделом диссертации использованы материалы следующих публикаций автора, в которых, согласно Положению о присуждении ученых степеней в МГУ, отражены основные результаты, положения и выводы исследования:

Ivanenkov Y. A., **Zagribelnyy B**., Aladinskiy V. Are we Opening the Door to a New Era of Medicinal Chemistry or Being Collapsed to a Chemical Singularity? // Journal of Medicinal Chemistry — 2019. — Т. 62. — №. 22. — С. 10026–10043.

Zhavoronkov A., Ivanenkov Y.A., Aliper A., Veselov M.S., Aladinskiy V.A., Aladinskaya A.V., Terentiev V.A., Polykovskiy D.A., Kuznetsov M.D., Asadulaev A., Volkov Y., Zholus A., Shayakhmetov R.R., Zhebrak A., Minaeva L.I., **Zagribelnyy B.**, Lee L.H., Soll R., Madge D., Xing L., Guo T., Aspuru-Guzik A. Deep learning enables rapid identification of potent DDR1 kinase inhibitors // Nature Biotechnology — 2019. — Т. 37. — № 9. — С. 1038–1040.

Патент US20220172802A1. Retrosynthesis systems and methods: опубл. 02.06.2022. / Konstantinov A., Putin E.O., **Zagribelnyy B**., Ivanenkov Y.A., Zhavoronkovs A.

Патент US20230154572A1. Retrosynthesis-related synthetic accessibility: опубл. 18.05.2023. / **Zagribelnyy B.**, Putin E.O., Fedorchenko S.A., Ivanenkov Y.A., Zavoronkovs A.

Патент WO2023078238A1. SARS-CoV-2 inhibitors for treating coronavirus infections: опубл. 11.05.2023. / Ding X., Peng J., Ren F., Ding X., **Zagribelnyy B**., Ivanenkov Y.A.

Ivanenkov Y. A., Polykovskiy D., Bezrukov D., **Zagribelnyy B**., Aladinskiy V., Kamya P., Aliper A., Ren F., Zhavoronkov A. Chemistry42: An AI-Driven Platform for Molecular Design and Optimization // Journal of Chemical Information and Modelling — 2023. — Т. 63. — № 3. — С. 695–701.

Ivanenkov Y., **Zagribelnyy B**., Malyshev A., Evteev S., Terentiev V., Kamya P., Bezrukov D., Aliper A., Ren F., Zhavoronkov A. The Hitchhiker's Guide to Deep Learning Driven Generative Chemistry // ACS Medicinal Chemistry Letters — 2023. — Т. 14. — №. 7. — С. 901–915.

Бондарев Н., **Загрибельный Б**., Федорченко С.А., Иваненков Я.А., Палюлин В.А. Моделирование синтетической доступности потенциальных лекарственных веществ, содержащих пятичленные ароматические гетероциклы // Известия Академии наук. Серия химическая. — 2025. — Т. 74. — № 6. — С. 1687–1703. (Переводная версия: Bondarev N., **Zagribelyy B.**, Fedorchenko S.A., Ivanenkov Ya. A., Palyulin V.A. Modeling of synthetic accessibility of potential drug molecules containing five-membered aromatic heterocycles // Russian Chemical Bulletin. — 2025. — Т. 74. — № 6. — С. 1687–1703.)

Sun J., Sun D., Yang Q., Wang D., Peng J., Guo H., Ding X., Chen Zh., Yuan B., Ivanenkov Y.A., Yuan J., **Zagribelnyy B**., He Y., Su J., Wang L., Tang J., Li Zh., Li R., Li T., Hu X., Liang X., Zhu A., Wei P., Fan Y., Liu S., Zheng J., Guan X., Aliper A., Yang M., Bezrukov D.S., Xie Zh., Terentiev V.A., Peng G., Polykovskiy D.A., Malyshev A.S., Malkov M.N., Zhu Q., Aspuru-Guzik A., Ding X., Cai X., Zhang Man, Zhao J., Zhong N., Ren F., Chen X., Zhavoronkov A., Zhao J. A novel, covalent broad-spectrum inhibitor targeting human coronavirus M^{pro}// Nature Communications — 2025. — Т. 16. — №. 4546.

модели GENTRL (см. рис. 19), дал толчок взрывному росту количества исследований в этой области. Несмотря на конструктивную критику результатов исследования [124,125], время показало, что общий тренд на использование инструментов генеративной химии и в частности ГО, заданный данным исследованием, является на данный момент доминирующим.

Рисунок 19. Архитектура эксперимента по генерации ингибиторов DDR1 киназы.

И хотя успех работы был вполне очевиден (*за 5 лет с момента публикации число цитирований статьи по версии платформы Scopus превысило отметку в 670 упоминаний*), одновременно очевидна была и потребность в создании интегрированной платформы, где не одна (как в случае GENTRL), а несколько генеративных моделей (например, десятки), будут в режиме реального времени взаимодействовать с модулями награды и фильтрами, непрерывно получая обратную связь от оценочных модулей. Причем число этих модулей может быть ограничено только целесообразностью использования и временем обработки молекулярных структур. В то время как генеративный подход при помощи модели GENTRL ограничивался использованием только четырёх модулей оценки и фильтрации:

1. Модуль структурных фильтров, основанных на экспертных знаниях в медицинской химии (МХФ, медхимические фильтры);
2. Самоорганизующиеся карты (СОК) Кохонена, обученные на известных киназных ингибиторах, DDR1 ингибиторах и химическом пространстве, соответствующем структурным трендам медицинской химии;
3. Модуль оценки соответствия фармакофорным моделям известных DDR1 ингибиторов;
4. Модуль оценки соответствия структурным трендам медицинской химии, MCE-18.

Стоит отметить, что оценка синтезируемости молекулярных структур в рамках эксперимента выполнялась вручную экспертами из компании WuXi AppTech, которые затем

и выполняли синтез отобранных структур. На тот момент адекватное задачам генеративной химии интегрированное решение для автоматизированной оценки синтезируемости ещё не было предложено.

Опираясь на опыт с моделью GENTRL, получив новые генеративные модели, а также запланировав новые оценивающие и фильтрационные модули, такие как модуль оценки синтезируемости, докинг-систему, модуль оценки соответствия форме темплатного лиганда и др., в конце 2019 года было принято решение интегрировать накопленные опыт и компетенции в платформу генеративной химии Chemistry42.

3.1.2 Верхнеуровневое описание архитектуры платформы генеративной химии Chemistry42

Верхнеуровневое описание платформы было представлено в публикации [126], подготовленной коллективом авторов группы компаний Insilico Medicine. Схематическое описание трехэтапного рабочего процесса для эксперимента с использованием платформы Chemistry42 представлено на рис. 20. На первом этапе пользователи загружают свои данные и настраивают платформу с желаемыми свойствами для генерируемых структур на защищенной и специфичной для компании версии программного обеспечения. Второй этап включает запуск платформы, где ансамбль из более чем 40 генеративных моделей функционирует параллельно для создания новых структур ─ этот этап называется фазой генерации. Разнообразные фильтры тщательно проверяют созданные молекулярные структуры в фазе генерации. Затем молекулярные структуры подвергаются анализу оценочными модулями, классифицируемыми на двумерные (2D) или трехмерные (3D) модули, которые динамически оценивают свойства созданных структур в соответствии с заранее определенными критериями. Дополнительные настраиваемые модули оценки (такие как прогноз ADME-свойств) также могут быть интегрированы в конвейер оценочных модулей для приоритизации созданных структур.

Рисунок 20. Верхнеуровневое описание платформы генеративной химии Chemistry42.

Эти модули составляют основу протокола генерации на базе многоагентного обучения с подкреплением (RL, *reinforcement learning*) в Chemistry42. Оценки созданных структур передаются обратно генеративным моделям для их усиления и направления процесса генерации к структурам с более высокими оценками — это называется фазой обучения. Финальный этап — это анализ. Созданные структуры автоматически ранжируются в соответствии с настраиваемыми метриками на основе их прогнозируемых свойств, включая синтетическую доступность, новизну, разнообразие и т. д. Платформа также предоставляет пользователям интерактивные инструменты для мониторинга производительности генеративных моделей.

Генеративные эксперименты создаются с использованием удобного веб-интерфейса платформы Chemistry42 и могут быть начаты с использованием методов разработки лекарств, основанных на лиганде или структуре, в зависимости от доступной информации о целевом объекте. Метод разработки лекарств, основанный на лиганде (LBDD), требует в качестве входных данных 2D или 3D структуру лиганда в виде файла .sdf, строки SMILES, или молекулу можно нарисовать непосредственно на платформе, используя удобную панель рисования. Также может быть добавлена фармакофорная гипотеза, которая создается вручную с помощью виджета или автоматически внутри платформы. В подходе к разработке лекарств, основанном на структуре биологической мишени (SBDD), структура белка-мишени, будь то в апо-формате или в комплексе с лигандом, должна быть загружена на платформу в виде подготовленного файла .pdb. Можно выбрать либо карман вокруг лиганда (сайт связывания лиганда), либо выбрать один из альтернативных карманов, указанных модулем Pocket Scanner. Как и в случае с LBDD, фармакофорная гипотеза может быть добавлена по при необходимости (см. рис. 21).

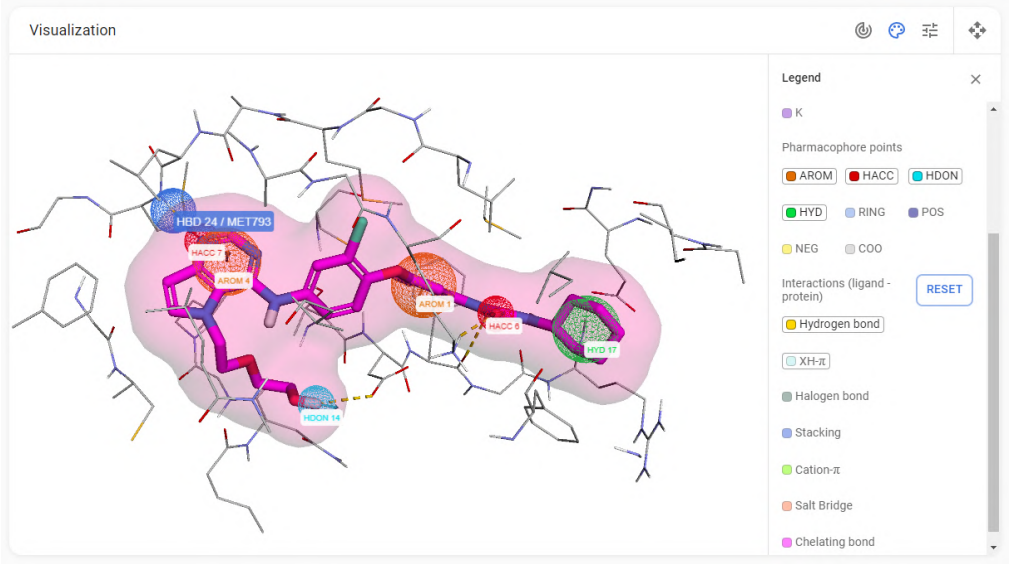


Рисунок 21. Пример фармакофорной гипотезы для генеративного эксперимента.

Для завершения настройки эксперимента генерации пользователь определяет допустимые диапазоны для множества свойств (например, физико-химических свойств и разнообразия) генерируемых структур. Пользователь может приоритизировать модули вознаграждения, регулируя их веса, и указать, насколько строгими должны быть модули, изменяя соответствующие пороги (см рис. 22).

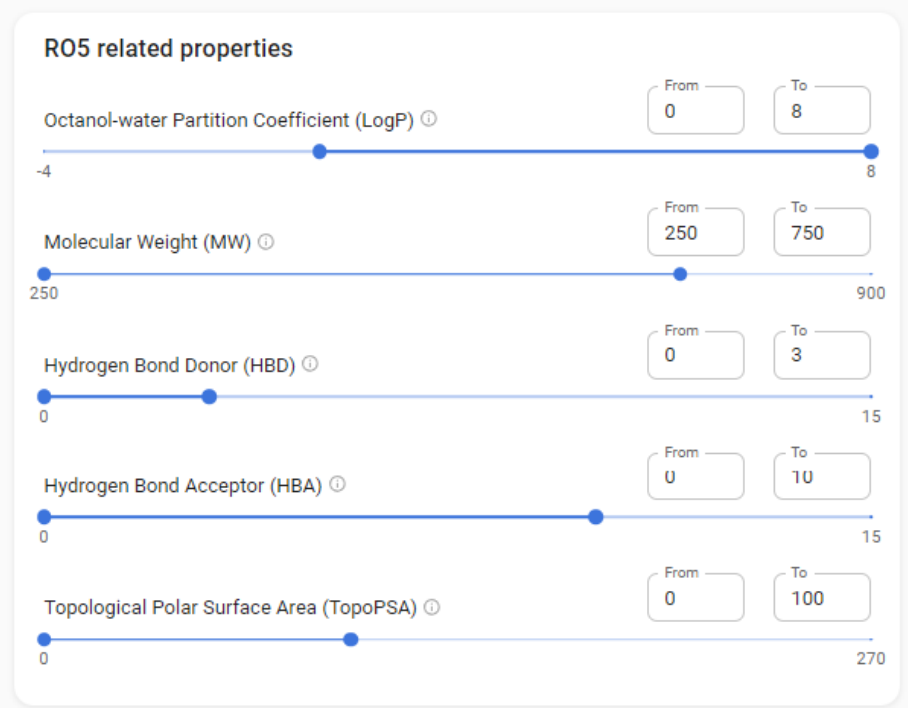


Рисунок 22. Интерфейс настройки физико-химических параметров, ассоциированных с правилами Липински [54].

В обоих подходах, LBDD и SBDD, расширенные параметры позволяют пользователю уточнять модули вознаграждения и выбирать, какие генеративные модели должны использоваться в эксперименте. Рабочие процессы по расширению хитов *(Hit expansion)*, оптимизации хитов и моделированию потенциальных лекарственных веществ на основе фрагментов (FBDD) также доступны на платформе с использованием функционала *Anchor points* (якорные точки, *англ*). С помощью *Anchor points* пользователи могут закреплять в 3D-пространстве определенные подструктуры или R-группы соединения-хита, в то время как остальная часть молекулы будет изменяться в ходе генеративного эксперимента (см рис. 23).

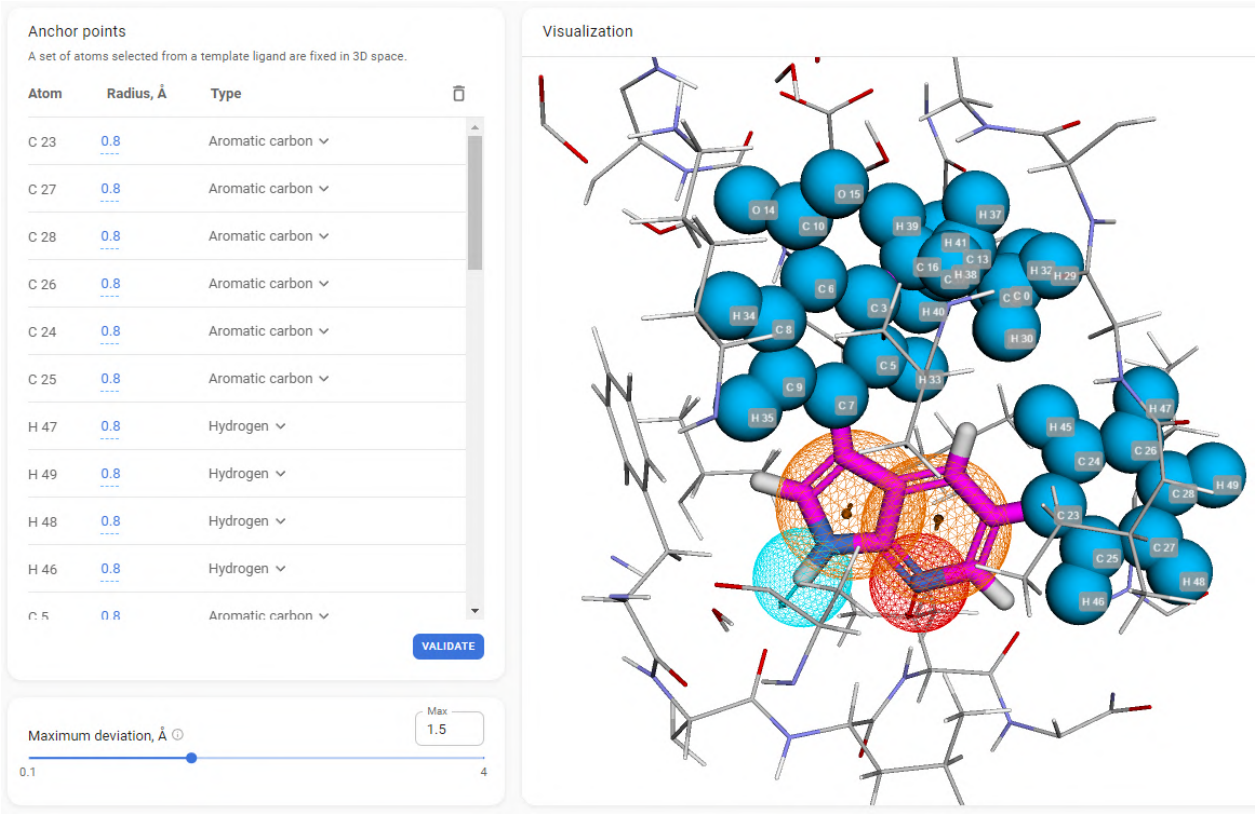


Рисунок 23. Пример использования якорных точек (голубые шарики) для фиксирования частей молекулы в трёхмерном пространстве.

Якорные точки также поддерживают возможность присваивать несколько разных атомных примитивов для каждого из закрепляемых атомов, позволяя изменять типы атомов в ходе генеративного эксперимента, не меняя их положение в 3D-пространстве с точностью до установленного отклонения. Например, пользователь может указать, хочет ли он видеть азот и/или углерод в определенной позиции ароматического кольца.

Генеративный конвейер в Chemistry42 включает асинхронный ансамбль проприетарных генеративных моделей. Эти тщательно отобранные алгоритмы имеют разнообразные архитектуры, реализующие различные стратегии. Платформа использует множество моделей машинного обучения и молекулярных представлений для различных генеративных сценариев, чтобы максимально использовать вклад каждой модели и повысить эффективность платформы. Например, некоторые модели сосредоточены на исследовании химического пространства, при одновременном улучшении обнаруженных структур. В текущей версии Chemistry42 представлено более 40 генеративных моделей, включая генеративные автокодировщики, генеративные состязательные сети, подходы на основе потоков, эволюционные алгоритмы, языковые модели и другие. Эти модели используют различные молекулярные представления: на основе строк, графов и 3D-представлений.

Взаимодействие множества генеративных моделей является важным аспектом. Поэтому, вместо того чтобы рассматривать эти алгоритмы как черные ящики, мы предоставляем глубокую предметно-ориентированную аналитику для понимания преимуществ и недостатков каждого подхода. Комбинируя различные передовые методы машинного обучения, Chemistry42 предоставляет разнообразные, высококачественные молекулярные структуры в течение нескольких часов. По мере их создания структуры динамически оцениваются с использованием модулей вознаграждения и оценки на платформе.

Модули вознаграждения и оценки, используемые в Chemistry42, могут быть двухмерными (2D) или трехмерными (3D). 2D-модули состоят из множества оценочных и фильтрационных модулей, включая медхимические фильтры (МХФ). В текущей версии Chemistry42 МХФ включают набор из более чем 460 проприетарных правил на основе SMARTS-подструктур, исключающих "плохие" структуры, то есть те, которые содержат PAINS-фрагменты [127] или функциональные группы, которые являются реакционноспособными, нестабильными или потенциально токсофорными. Функция *Medicinal Chemistry Evolution* (MCE-18) представляет собой уникальный молекулярный дескриптор, который оценивает структуры на предмет их соответствия структурным трендах медицинской химии. Другие 2D модули включают правило пяти Липински [54,127], оценку подобия лекарствам (drug-likeness) и дескрипторы взвешенной доли атомных типов, а также фильтр на основе правил, который исключает структуры с несбалансированным количеством гетероатомов и ароматических атомов. Оценка новизны (*Novelty*) основана на анализе 2D-сходства между созданными структурами и референсным набором данных (который может быть настроен). Синтетическая доступность (СД) созданных структур моделируется при помощи метода ReRSA, подробно описанного в разделе 2.2. Отслеживание разнообразия (*Diversity*) созданных структур предоставляет средство для понимания того, насколько структурно разнообразны созданные молекулы, на основе количества сгенерированных хемотипов после кластеризации.

Привилегированные фрагменты (ПФ) — это определенные структурные мотивы, которые способствуют активности в отношении мишени или класса мишеней. Функциональность ПФ наиболее полезна в двух типах рабочих процессов генеративного дизайна. Первый включает определение 2D-ПФ подструктур, которые будут присутствовать во всех созданных структурах без предварительно определенного позиционирования в 3D пространстве. Это полезно, если у пользователя есть только апо-структура биологической мишени без описанных лигандов. Например, если мишень — это апо-структура плохо изученной протеиновой киназы, то 2D-ПФ соединений с петлевыми узлами могут быть

использованы в генеративных экспериментах для навигации по известному химическому пространству. Второй рабочий процесс включает использование **якорных точек** (англ., *Anchor Points*) — по сути 3D-привилегированных фрагментов. Здесь присутствие интересующей подструктуры важно либо в комплексе белок–лиганд (режим SBDD), либо в 3D конформации лиганда (режим LBDD). Модуль классификации самоорганизующихся карт (СОК) Кохонена (общая карта СОК 100×100) используется для направления генерации молекулярных структур в химическое пространство, соответствующее указанному классу мишеней. Поскольку общая СОК содержит нейроны с классификационной способностью ниже заданного порога для выбранной категории молекул, все эталонные молекулы из таких нейронов собираются и затем подвергаются автоматически сгенерированным ZOOM картам адаптированного размера для достижения надежной классификационной точности. Набор данных, используемый для обучения модуля классификации СОК, и ZOOM карты называются набором данных Hierarchical Active Molecules (HAM). Набор данных HAM состоит из данных о более 800 тыс. молекул, подтвержденных экспериментально с IC_{50} 10 μM или меньше. Модуль морфинга структуры включает два компонента: усилитель метаболической стабильности, основанный на правилах, который решает проблемы метаболической нестабильности, вызванной потенциальными сайтами метаболизма в созданных структурах, и модуль биоизостеров, который выполняет биоизостерические/изостерические трансформации.

После оценки созданных структур с помощью 2D модулей, для дальнейшей оценки используются несколько 3D модулей. Первым из них является модуль ConfGen, который создает конформационный ансамбль для каждой созданной структуры. Модуль ConfGen генерирует конформационные ансамбли с использованием набора внутренних правил и заранее определенных геометрий подструктур на основе данных рентгеноструктурного анализа малых молекул, кристаллизованных с белками, за которым следует минимизация энергии с использованием проприетарного силового поля Insilico. Для ранжирования молекулярных структур по внутренней жесткости используется оценка гибкости (FLEX score). После создания конформационных ансамблей модуль 3D-дескрипторов оценивает 3D-сходство между созданными структурами и эталонной молекулой (входной лиганд) с использованием набора рассчитанных 3D-дескрипторов. Затем модуль фармакофора оценивает, соответствуют ли какие-либо из созданных конформаций заданной гипотезе фармакофора, включая все важные точки связывания, расстояния, углы и допустимые отклонения. Если в генерации используется модуль Anchor Points, он проверяет, присутствуют ли в созданной структуре заданные пользователем 3D подструктуры в правильном положении

и конформации. Модуль Shape Similarity оценивает 3D-сходство формы с эталонной молекулой, используя взвешенные гауссовы функции. Последний модуль сосредотачивается на позиционировании и оценке созданных структур, чтобы оценить, насколько хорошо они вписываются в выбранное место связывания (модуль кармана), и приближает силу связывания с помощью оценки взаимодействия лиганд-карман (PLI score, *pocket-ligand interactions score*). Оценка PLI была обучена на уточненном наборе данных PDBBind v2020 [128] (использовались данные как по K_i, так и по K_d). Оценка учитывает водородные связи, π-стекинговые, π-катионные, XH-π и гидрофобные взаимодействия, а также солевые мостики и хелатные связи. Единицы оценки PLI — ккал/моль, причем чем более отрицательное значение, тем лучше оценка.

Пользователь может указать, как долго он хочет запускать генеративный эксперимент. В большинстве случаев наблюдается сходимость через 72 часа. Во время генеративного эксперимента производительность каждой генеративной модели отслеживается и записывается. Это позволяет пользователю следить за ходом своих экспериментов в реальном времени с начала до завершения. Созданные структуры автоматически оцениваются и ранжируются в соответствии с метриками, встроенными в модули платформы. Все соответствующие данные, включая оценки, молекулярные структуры и производительность генеративных моделей, хранятся и доступны на странице результатов платформы. После завершения генеративного эксперимента результаты могут быть проанализированы с помощью интерактивного интерфейса.

Средний пользователь может получить ценные результаты (1–5 новых молекул для синтеза) даже при первом запуске генерации для SBDD, LBDD и других различных рабочих процессов. Полученные результаты и их последующий анализ могут помочь настроить второй запуск более конкретным образом для получения более ожидаемых результатов. Обычно при втором запуске пользователь может использовать некоторые идеи, предложенные платформой при первом запуске (например, добавить некоторые привилегированные структуры при втором запуске). Опытный пользователь, который глубоко знаком с функциональностью и результатами работы платформы, может настроить платформу и получить ожидаемые результаты (10–20 новых молекул для синтеза) даже при первом запуске нового проекта.

Платформа предоставляет пользователю хорошо документированное онлайн-руководство, описывающее примеры модельных экспериментов для сценариев работы, характерных при использовании классических методов дизайна потенциальных лекарственных веществ, пошаговые инструкции для этих модельных экспериментов и

примеры результатов в виде sdf файлов и их анализ медицинскими и вычислительными химиками. Эти модельные эксперименты могут помочь новичкам выбрать более подходящую стратегию использования платформы для их собственных проектных целей.

3.1.3 Модельные эксперименты в рамках платформы Chemistry42

Создание модельных экспериментов стало важнейшей вехой в разработке платформы генеративной химии Chemistry42 по нескольким причинам:

1. Во-первых, модельные эксперименты призваны проиллюстрировать функциональность платформы с точки зрения практической медицинской химии и моделирования потенциальных лекарственных веществ. Это важно как с точки зрения бизнес-процессов и позиционирования платформы для потенциальных клиентов, так и с точки зрения внутренней разработки.
2. Во-вторых, упомянутый выше интерес внутренних разработчиков к наличию модельных экспериментов обуславливается тем, что благодаря стабильной системе подобных экспериментов разработчики платформы могут отслеживать прогресс в развитии и улучшении компонентов платформы. Так, например, при добавлении новых генеративных моделей в общий ансамбль можно проследить на основе модельных экспериментов какой вклад новая модель вносит в общий результат работы платформы и на основании метрик принять решение о дальнейшем внедрении новой генеративной модели в общий пул моделей по умолчанию.
3. В-третьих, наличие модельных экспериментов крайне полезно для обучения новых клиентов работе на платформе. Вместо изучения “сухого” руководства пользователя клиенту предлагается пройти процесс обучения через запуск модельных экспериментов и сравнить полученные результаты с эталонными, которые, в свою очередь, доступны к загрузке с платформы.

В дизайне модельных экспериментов мы руководствовались реальной практикой современной медицинской химии, которая иллюстрируется примерами компьютеризированного дизайна малых лекарственных молекул из ведущих профильных журналов (в первую очередь *Journal of Medicinal Chemistry* [129]). Создание модельных экспериментов согласовано с классификацией базовых сценариев дизайна потенциальных лекарственных веществ, приведенных в таблице 1 (см. разд. 1.1.2). В настоящем диссертационном исследовании рассмотрим шесть модельных экспериментов, приведенных в табл. 13.

Таблица 13. Модельные эксперименты платформы генеративной химии Chemistry42

#	Стратегия дизайна	Модельный эксперимент
1	Виртуальный скрининг	Виртуальный скрининг ингибиторов папаин-подобной протеазы коронавируса SARS-CoV-2
2	*De novo* дизайн	Генеративный *de novo* дизайн ингибиторов Jak3 киназы
3	*Hit-expansion*	Генеративный дизайн аналогов соединения-хита протеазы USP7
4	*Scaffold-hopping*	Генеративный *scaffold-hopping* дизайн ингибиторов CAMKK2 киназы
5	Дизайн R-групп	Генеративный дизайн заместителей ингибитора MPS1 киназы
6	FBDD	Генеративный дизайн ингибиторов главной протеазы коронавируса SARS-CoV-2 на основе знаний о связывании малого фрагмента

Отметим, что все шесть перечисленных модельных экспериментов предлагается проводить в смешанной SBDD/LBDD парадигме, то есть в рамках моделирования потенциальных лекарственных веществ, как на основе знаний о структуре мишени, так и на основе известных лигандов, причем в качестве лиганда предлагается брать тот, что находится в со-кристаллизованной с мишенью форме и представлен таким образом в виде pdb файла в банке данных PDB.

3.1.3.1 Виртуальный скрининг ингибиторов папаин-подобной протеазы коронавируса SARS-CoV-2

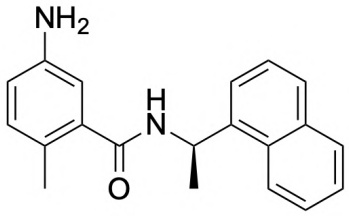
Несмотря на то, что виртуальный скрининг не представляет собой тип эксперимента, согласующийся с понятием генеративной химии, но даже и противопоставляется концепту генеративной химии, тем не менее, знакомство с платформой генеративной химии Chemistry42 предлагается начать именно с эксперимента по виртуальному скринингу. Делается это по причине того, что фактически без генеративных моделей платформа Chemistry42 представляет собой классический движок по проведению виртуального скрининга. Поскольку клиенты зачастую хорошо представляют то, как работает виртуальный скрининг, и какие ожидания у них есть относительно работы подобных движков, то такое первичное знакомство с платформой через привычный эксперимент представляется логичным. Более того, именно через эксперимент по типу виртуального скрининга клиент может напрямую удостовериться

в том, что функционал платформы работает корректно и соответствует физической картине о том, как, например, низкомолекулярные агенты взаимодействуют с макромолекулами.

Преамбула эксперимента

Папаин-подобная протеаза (PLpro) коронавируса SARS-CoV-2 является ключевым ферментом коронавируса, который способствует распространению вируса через обработку вирусных полипротеинов, что приводит к образованию функционального комплекса репликазы [130]. PLpro SARS-CoV-2 считается одной из наиболее перспективных мишеней для разработки новых противовирусных препаратов для лечения COVID-19. Виртуальный скрининг существующих ингибиторов протеазы может помочь выявить потенциальные отправные точки для разработки новых ингибиторов PLpro SARS-CoV-2. В настоящее время производные аминонафталина считаются наиболее перспективными нековалентными ингибиторами PLpro с субмикромолярной активностью. В этом модельном эксперименте по виртуальному скринингу мы будем использовать кристаллическую структуру папаиноподобной протеазы в комплексе с GRL0617 — примером производного аминонафталина (см табл. 14). Это упражнение включает в себя подготовку входных файлов и использование внешних наборов данных для виртуального скрининга.

Таблица 14. Входные данные и ключевые 3D модули для модельного эксперимента №1

Исходный PDB файл	**Темплатный Лиганд**
7CMD [131]	 **GRL0617**
Ключевые 3D модули	
PLI Score Фармакофорный модуль Модуль оценки подобия формы	

Ход эксперимента

Пользователю предлагается использовать набор данных об ингибиторах различных протеаз, который будет использован в этом модельном эксперименте для скрининга, был получен из базы данных ChEMBL [116]. Набор данных состоит из соединений с pchembl_value ≥ 5.0. Значение pchembl_value представляет собой универсальную шкалу

активности для молекул в ChEMBL, используемую как –lg (Параметр) для следующих параметров: 'Potency', 'IC_{50}', 'K_i', 'EC_{50}', 'K_d'. Например, в терминах IC_{50}, значение pchembl_value 5.0 соответствует значению IC_{50} в 10μM. Набор данных также включает соединения с уровнями достоверности биологических тестов ≥ 9 [132] и типом теста: B (*binding*, англ. связывание), F (*functional*, англ. функциональный). Дубликаты и кофакторы солей были удалены в процессе стандартизации. Были применены мягкие медхимические фильтры (МХФ) для исключения молекул, не обладающих подобием лекарствам (например, металлов, поликонденсированной ароматики, хлораминов, свободных радикалов, гидразинов, изонитрилов, нитрозосоединений), структур, содержащих циклы с более чем 8 атомами, и полипептидов ($n \geq 4$). Итоговый набор данных содержит 54 205 уникальных структур протеазных ингибиторов и предоставляется для загрузки пользователю.

В качестве движущей силы отбора в ходе виртуального скрининга предлагается использовать трёхточечную фармакофорную гипотезу (две ароматические фармакофорные точки AROM и один донор водородной связи HDON), сформированную на основе со-кристаллизованного лиганда **GRL0617** и анализа важности взаимодействий с аминокислотными остатками (см. рис. 24).

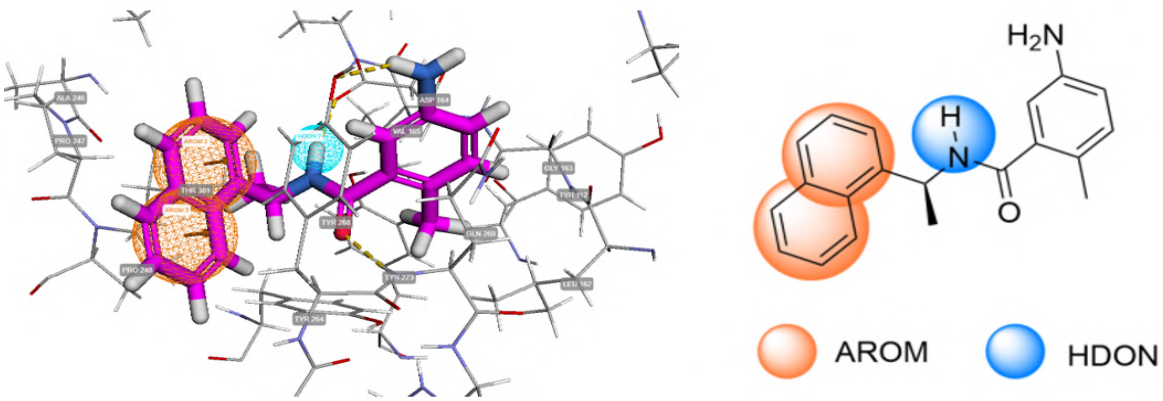


Рисунок 24. Фармакофорная гипотеза для модельного эксперимента №1.

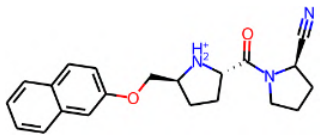
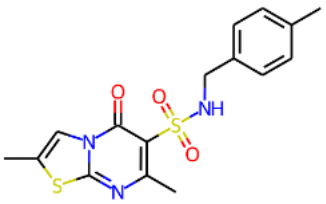
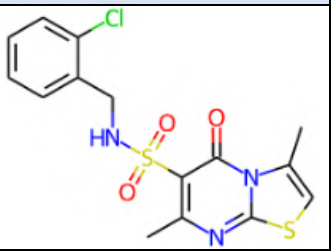
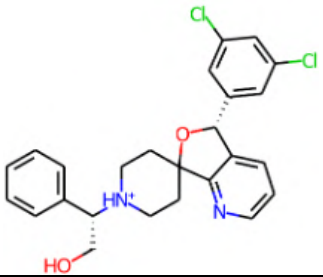
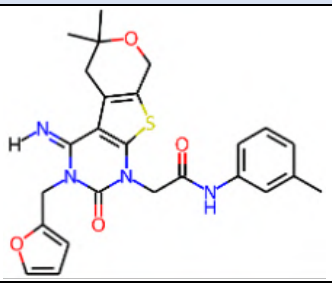
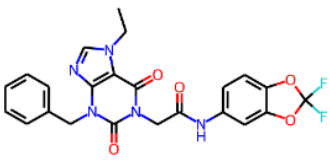
Помимо фармакофорного модуля вклад в итоговое значение функции награды будут вносить PLI Score, модуль оценки подобия форме темплатного лиганда (далее Shape модуль) и СОК Кохонена, сфокусированные на цистеиновых протеазных ингибиторах.

Эталонные результаты

По итогам модельного эксперимента пользователю доступна таблица с результатами виртуального скрининга, в которой оцененные молекулярные структуры отсортированы в порядке уменьшения Значения функции награды (Reward), которая представляет собой

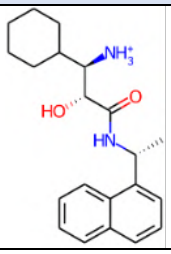
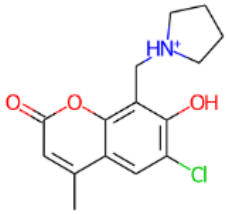
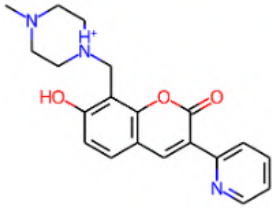
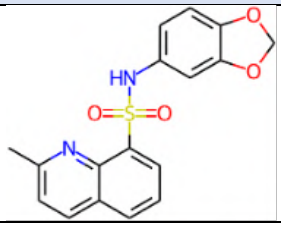
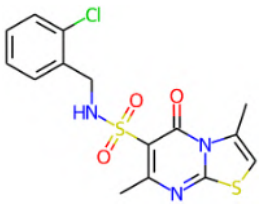
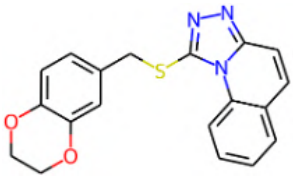
агрегированное и безразмерное значение для оценок выдаваемых, как всеми оценочными модулями, так и всеми фильтрационными модулями. Однако далеко не всегда пользователя может интересовать такая агрегированная оценка, в то время как он желает уделить большее внимание интерпретируемым с физической точки зрения модулям и генерируемым ими значениям оценок. Так, например, главный интерес представляет значение PLI Score, которое является аналогом ЗОФ (значение оценочной функции) докинга на платформе Chemistry42. Также физически интерпретируемыми являются фармакофорный модуль и нахождение молекулярных структур с высокими значениями фармакофорного подобия. Несколько менее связанным с активностью является фактор подобия форме темплатному лиганду, однако в комбинации с высокими значениями фармакофорного подобия этот физически интерпретируемый фактор может сыграть положительную роль в отборе молекулярных структур на биологические тестирования. Исходя из вышесказанного, в качестве примера, мы предлагаем пользователю обратить внимание на 3 лучших структуры по PLI Score, оценке фармакофорного подобия и оценке подобия формы (см. табл. 15).

Таблица 15. Результаты модельного эксперимента №1

3 лучшие структуры по значениям функции награды (Reward)			
Структура			
ID	INS-009923	INS-049219	INS-060666
Reward	**3.53**	**3.50**	**3.47**
ReRSA	2.46	5.91	5.84
MCE-18	70.67	40.00	40.00
3 лучшие структуры по PLI Score			
Структура			
ID	INS-036683	INS-058085	INS-043149
PLI Score	**-10.54**	**-10.45**	**-10.38**
ReRSA	6.61	5.07	2.64
MCE-18	117.82	65.45	62.14

Продолжение на следующей странице

Продолжение таблицы 15

3 лучшие структуры по оценке фармакофорного подобия (Ph4 Score)			
Структура			
ID	INS-039265	INS-059638	INS-051552
Ph4 Score	0.95	0.94	0.91
ReRSA	2.07	6.08	4.47
MCE-18	57.58	40.48	48.46
3 лучшие структуры по оценке подобия формы (Shape score)			
Структура			
ID	INS-060203	INS-060666	INS-046994
Shape score	0.83	0.79	0.79
ReRSA	2.57	5.84	5.54
MCE-18	46.32	40.00	50.18

При помощи встроенного молекулярного визуализатора пользователь может удостовериться в достоверности предлагаемых значений оценочных модулей для произведенных генеративными алгоритмами молекулярных структур, как в трехмерном пространстве с доступным наложением относительно темплатного лиганда, так и при анализе двумерных молекулярных структур.

Предполагается, что, руководствуясь вышеупомянутыми соображениями и исходя из доступных ресурсов, пользователь сможет отобрать на платформе из результатов виртуального скрининга те молекулярные структуры, которые он хотел бы протестировать в биологической лаборатории.

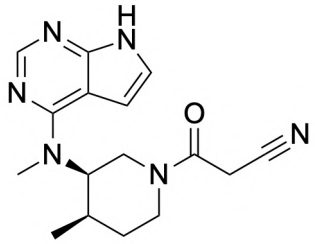
3.1.3.2 Генеративный de novo дизайн ингибиторов Jak3 киназы

Преамбула эксперимента

Киназа Jak3 является важной тирозинкиназой, участвующей в сигнальном пути JAK-STAT и необходима для развития иммунных клеток. Ингибирование Jak3 может приводить к терапевтическому эффекту в лечения аутоиммунных и тяжелых воспалительных заболеваний. Например, одобренный FDA препарат тофацитиниб (селективный ингибитор Jak3), разработанный компанией Pfizer, используется для лечения ревматоидного артрита и

язвенного колита (см. табл. 16). В этом исследовании на примере мы будем генерировать соединения, которые являются новыми ингибиторами Jak3 и соответствуют правилу пяти Липински, используя общедоступные данные [133]:

Таблица 16. Входные данные и ключевые 3D модули для модельного эксперимента №2

Исходный PDB файл	**Темплатный Лиганд**
3LXK [134]	 Тофацитиниб
Ключевые 3D модули	
PLI Score Фармакофорный модуль Модуль оценки подобия формы	

Ход эксперимента

Пользователю предлагается, пользуясь инструментарием платформы создать четырёхточечную фармакофорную гипотезу, соответствующую классическому фармакофору киназных ингибиторов: акцептор (HACC) и донор (HDON) водородных связей, две ароматических фармакофорных точки в hinge регионе сайта связывания АТФ (аденозинтрифосфата) (см. рис. 25). Наряду с фармакофорным модулем движущую силу генеративного эксперимента будут представлять модуль оценки взаимодействий белок-лиганд (*PLI*) и модуль оценки подобия формы (*Shape*), который будет сопоставлять форму генерируемых молекулярных структур с трёхмерной конформацией тофацитиниба. Помимо этого, предлагается установить настройки физико-химических параметров таким образом, чтобы они приводили генерацию в физико-химическое пространство, ограниченное правилами Липински (см. рис. 22) и СОК Кохонена, сфокусированные на известных ингибиторах Jak3.

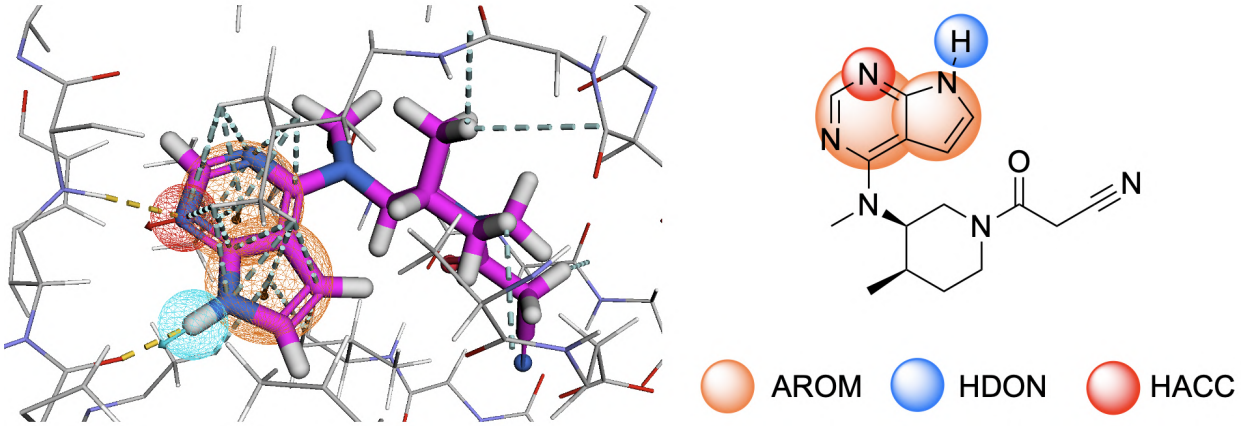


Рисунок 25. Фармакофорная гипотеза для модельного эксперимента №2.

Помимо фармакофорных точек, ключевые взаимодействия между известными лигандами и их мишенью могут быть установлены в виде **обязательных точек** в сайте связывания (см. рис. 26). В текущей версии платформы можно выбрать до одной обязательной точки связывания. Эта обязательная точка, представляющая собой атом на аминокислотном остатке, гарантирует, что взаимодействие (на уровне молекулярного докинга) произойдет между сгенерированными соединениями и мишенью. В текущем модельном эксперименте *hinge*-регион имеет решающее значение для связывания с киназами, и пользователь может выбрать акцептор (HBA) или донор (HBD) со стороны белка в этой области сайта связывания АТФ. В этом конкретном случае обязательной точкой связывания может быть либо донор N-H LEU905, либо акцептор C=O GLU903 загруженной кристаллической структуры Jak3, связанной с лигандом (см. рис. 26).

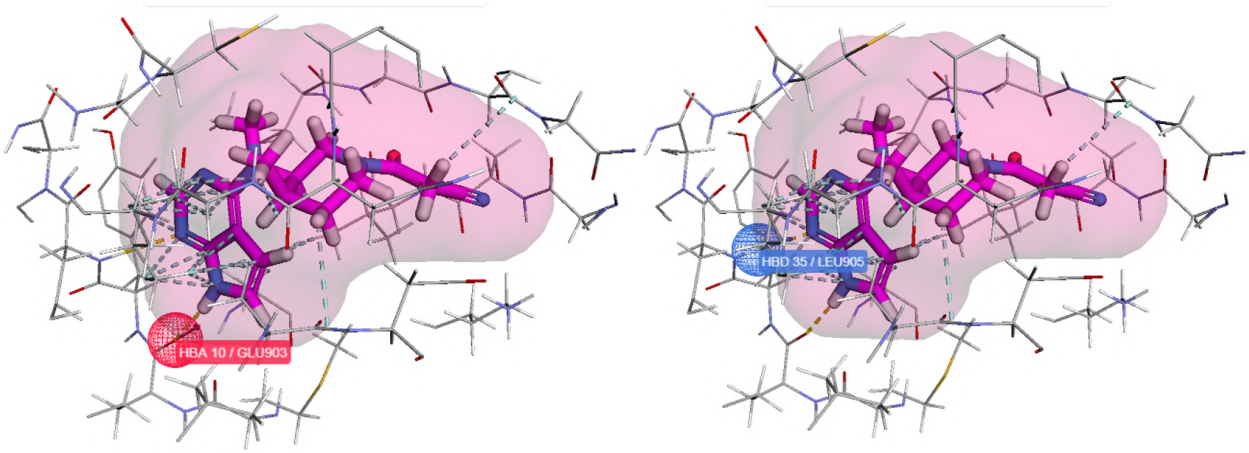


Рисунок 26. Выбор обязательной точки для модельного эксперимента №2.

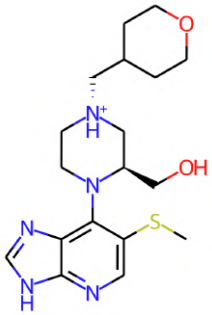
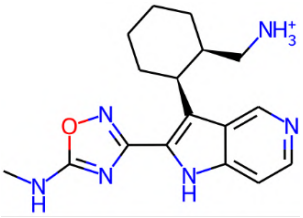
Эталонные результаты

По итогам генеративного эксперимента пользователю предлагается обратиться к таблице результатов, содержащей информацию о кластеризованном наборе молекулярных структур (см. табл. 17). Кластеризованный набор молекулярных структур отличается от

общего набора результирующих молекулярных структур тем, что содержит лишь лучших (по награде, Reward) представителей кластеров. Интерес данный набор может представлять, если пользователь желает сэкономить время на ознакомление с генерацией (всего в среднем на выходе из этого эксперимента генерируется около 2500 молекулярных структур) и быстрее перейти к отбору молекулярных структур на синтез.

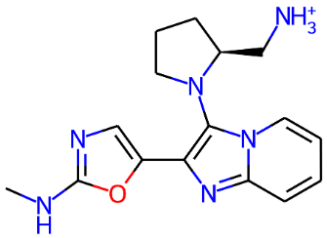
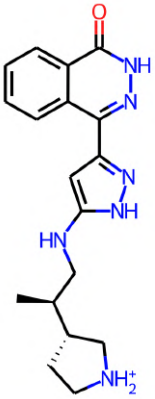
Как было ранее сказано (см. *Эталонные результаты* в разд. 3.1.3.1) существует множество метрик, позволяющих провести отсев молекулярных структур: это и агрегированная награда (Reward), и более физически интерпретируемые PLI Score, оценка фармакофорного подобия (Ph4 score) и подобия формы (Shape score). Также при размере фармакофорной гипотезы более 3 точек интерес может представлять метрика количества удовлетворенных точек фармакофорной гипотезы (*Ph4 points matched*). Пользователя могут, например, интересовать только те молекулы, которые “попали” во все точки фармакофорной гипотезы.

Таблица 17. Результаты модельного эксперимента №2

ID	**X1975-4178-0159**	**X1975-4178-0157**
Структура		
PLI Score	-8.62	-9.38
Shape score	0.94	0.74
Ph4 points matched	4/4	3/4
ReRSA	2.47	5.23
MCE-18	72.00	69.60
Reward	3.87	3.66

Продолжение на следующей странице

Продолжение таблицы 17

ID	X1975-4178-0277	X1975-4178-1752
Структура		
PLI Score	-9.24	-9.29
Shape score	0.75	0.67
Ph4 points matched	3/4	3/4
ReRSA	3.11	2.35
MCE-18	67.27	66.40
Reward	3.60	3.21

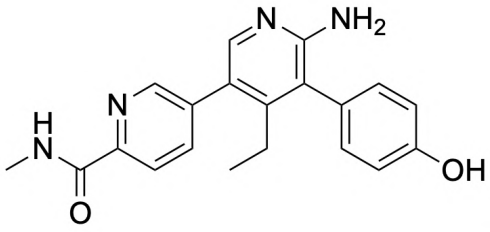
3.1.3.3 Генеративный дизайн аналогов соединения-хита протеазы USP7

Преамбула эксперимента

Убиквитин-специфическая протеаза 7 (USP7, *Ubiquitin-specific protease 7*) также известна как герпесвирус-ассоциированная убиквитин-специфическая протеаза (HAUSP). В течение последнего десятилетия обширный исследовательский интерес к USP7 выявил ее роль в различных клеточных путях, включая регуляторы вирусных белков, иммунный ответ, восстановление повреждений ДНК, контроль клеточного цикла и апоптоз. Было показано, что аномальная экспрессия USP7 при различных видах рака регулирует динамику сети p53-MDM2 и способствует возникновению и прогрессированию опухолей, что делает ее привлекательной целью с терапевтической точки зрения. Разработка ингибитора USP7, который бы усиливал эндогенное убиквитинирование MDM2 и стабилизировал p53, сдерживается отсутствием кристаллографических структур протезы в комплексе с низкомолекулярными ингибиторами и неспособностью обеспечить селективность по сравнению с другими высокоактивными деубиквитиназами (USP5, USP47). Тем не менее, относительно недавно была опубликована структура USP7 в комплексе с соединением-хитом **GNE6776** [135], обладающим посредственным значением IC_{50} (1.34μM), что, однако, может являться начальной точкой в расширении химического пространства вокруг соединения-хита и дальнейшей оптимизации хемотипа.

В настоящем модельном эксперименте пользователю предлагается провести эксперимент по генерации аналогов **GNE6776** (см. табл. 18) с целью расширения химического пространства вокруг соединения-хита (*Hit expansion*).

Таблица 18. Входные данные и ключевые 3D модули для модельного эксперимента №3

Исходный PDB файл	**Темплатный Лиганд**
5UQX [136]	 **GNE6776**
Ключевые 3D модули	
PLI Score Якорные точки Фармакофорный модуль Модуль оценки подобия формы	

Ход эксперимента

В первую очередь пользователю предлагается выбрать якорные точки для спецификации фрагмента, который должен быть закреплен в трёхмерном пространстве. Исходя из анализа структурной информации связывание фенольного гидроксила **GNE6776** с остатком гистидина, является ключевым фактором, приносящим положительный вклад в активность **GNE6776**. Таким образом, фенольный фрагмент **GNE6776** будет закрепляться при помощи якорных точек (см. рис. 27). Помимо этого, авторы исследования утверждают, что положительный вклад в связывание вносят взаимодействия с остатком аспарагиновой кислоты Arg305 (водородная связь с NH-группой амидного фрагмента **GNE6776**) и тирозина Tyr348 (обильные XH-π контакты с аминопиридиновым кольцом). Два последних указанных взаимодействия предлагается выразить в виде фармакофорных точек (HDON и AROM соответственно, см. рис. 27).

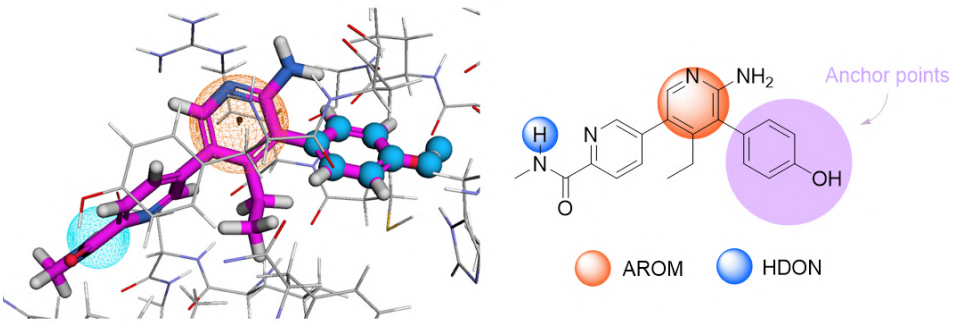


Рисунок 27. Фармакофорная гипотеза и якорные точки для модельного эксперимента №3.

Таким образом, помимо основных ранее упомянутых 3D модулей, оказывающих решающий вклад в значение функции награды (PLI, фармакофорный модуль и модуль подобия формы), в этом модельном эксперименте важную роль будет играть модуль якорных точек.

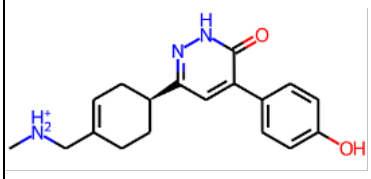
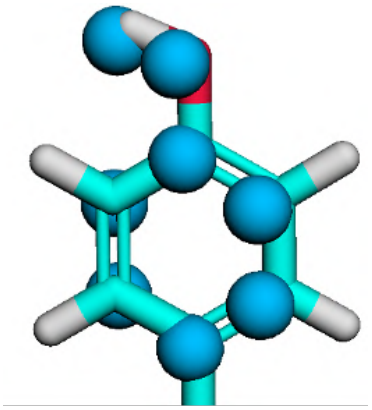
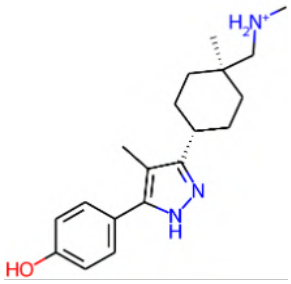
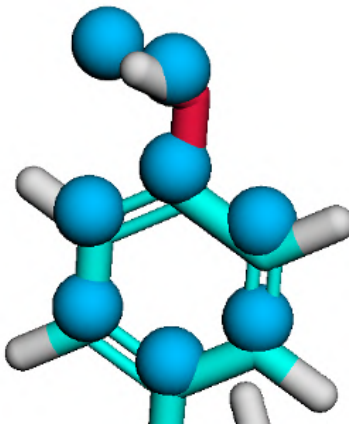
Эталонные результаты

Процедура анализа результатов для модельного эксперимента №3 похожа на ту, что была в случае первых двух (см. разд. 3.1.3.1 и 3.1.3.2). Однако момент, который следует учитывать для этого типа генерации, — это оценка наложения на якорные точки (*Anchor points score*), которая показывает, насколько удовлетворительно выполнено выравнивание сгенерированной структуры относительно якорных точек из темплатного лиганда, где 0.01 — это худшая оценка, которую можно наблюдать, а 1.0 — идеальное соответствие между 3D-привилегированным фрагментом в сгенерированной структуре и привилегированным фрагментом в темплатном лиганде. Обычно удовлетворительные выравнивания наблюдаются начиная со значений *Anchor points score* 0.4–0.45. Для иллюстрации эффективности наложения модулем якорных точек в таблице 19 приведена 3D-визуализация наложения фенольных фрагментов сгенерированных структур из кластеризованного набора на соответствующие якорные точки темплатного лиганда (голубые шары).

Как мы писали выше (см. *Эталонные результаты* в разд. 3.1.3.2), хорошим критерием для отбора молекулярных структур является полное удовлетворение фармакофорной гипотезы с точки зрения числа удовлетворенных точек гипотезы (Ph4 points matched). Последний этап отбора может быть сделан на основе PLI Score. Следует отметить, что этот порядок фильтрации не является единственным верным и может быть изменен пользователем платформы на основе его личного понимания важности признака/оценки.

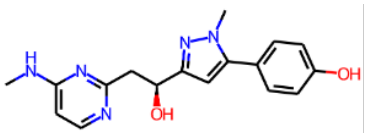
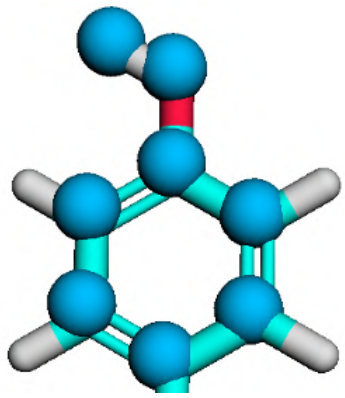
Описанный модельный эксперимент может быть полезен для решения задач расширения химического пространства (*Hit expansion*). Так, например, в среднем, удается сгенерировать около 2 000 молекулярных структур аналогов **GNE6776**.

Таблица 19. Результаты модельного эксперимента №3

3 лучшие структуры по функции награды (Reward)		
ID	**Структура**	**Наложение на якорные точки**
X1975-8222-0001		
PLI Score	-9.27	
Ph4 Score	0.95	
Ph4 points matched	2/2	
Anchor points score	0.88	
Shape score	0.84	
ReRSA	3.77	
MCE-18	53.33	
Reward	3.35	
X1975-8222-0031		
PLI Score	-8.75	
Ph4 Score	0.95	
Ph4 points matched	2/2	
Anchor points score	0.91	
Shape score	0.81	
ReRSA	3.76	
MCE-18	43.45	
Reward	3.27	

Продолжение на следующей странице

Продолжение таблицы 19

ID	Структура	Наложение на якорные точки
X1975-8222-0065		
PLI Score	-8.48	
Ph4 Score	0.95	
Ph4 points matched	2/2	
Anchor points score	0.98	
Shape score	0.81	
ReRSA	3.04	
MCE-18	34.00	
Reward	3.25	

3.1.3.4 Генеративный scaffold-hopping дизайн ингибиторов CAMKK2 киназы

Преамбула эксперимента

Кальций (Ca^{2+})/кальмодулин-зависимая протеинкиназа киназа 2 (CAMKK2) — это серин/треониновая киназа, которая является одним из связывающих кальмодулин (CaM) белков семейства CaMK. После активации CAMKK2 фосфорилирует и активирует свои субстраты, включая CAMK1, CAMK4, АМФ-активируемую протеинкиназу (AMPK) и, в некоторых случаях, AKT. Эта передача сигнала приводит к регуляции многих важных физиологических и патологических процессов. Аберрантная активация и сверхэкспрессия CAMKK2 были связаны с несколькими типами рака, включая рак простаты, молочной железы, яичников, желудка и печени.

В недавно опубликованном материале описывается выполнение *scaffold hopping* дизайна с целью расширить химическое пространство вокруг соединения-хита, ингибитора CAMKK2 (**GSK650394**) [137]. Авторы исследования демонстрируют, что путем замены исходного азаиндоального скаффолда на другие конденсированные и неконденсированные системы (см. рис. 28), удается найти около двух десятков новых соединений-хитов с похожим уровнем активности.

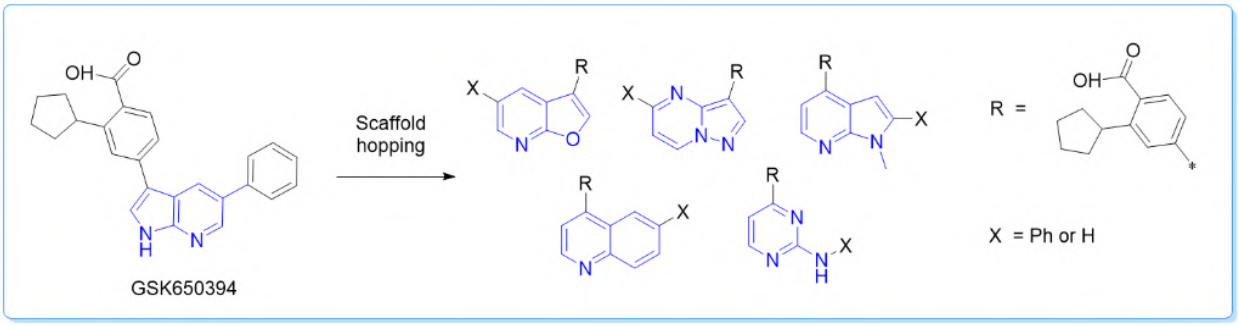
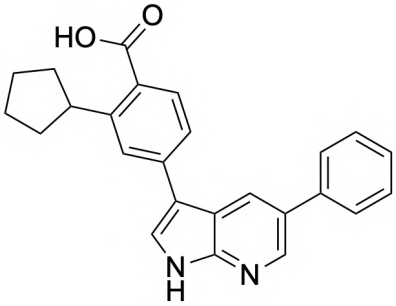


Рисунок 28. Примеры выполнения *scaffold hopping* дизайна для ингибитора CAMKK2 [137].

В рамках модельного эксперимента №4 пользователям предлагается произвести на платформе похожий *in silico* эксперимент, используя данные о комплексе **GSK650394** с CAMKK2 (см. табл. 20), в целях убедиться в том, что платформа Chemistry42 позволяет найти как ранее опубликованные примеры скаффолдов, так и значительное количество тех, которые не были приведены в оригинальной статье и представляют интерес для дальнейших исследований (синтез и биологические тестирования).

Таблица 20. Входные данные и ключевые 3D модули для модельного эксперимента №4

Исходный PDB файл	**Темплатный Лиганд**
6BKU [138]	 **GSK650394**
Ключевые 3D модули	
PLI Score Якорные точки Фармакофорный модуль Модуль оценки подобия формы	

Ход эксперимента

Как и в предыдущем модельном эксперименте пользователю предлагается закрепить ту часть молекулы, которая не должна изменять при помощи якорных точек. В данном случае, как и в оригинальной статье, изменению не будут подвергаться мотив о-циклопентилбензойной кислоты и незамещенный фенил (см. рис. 29). В то же время, оставшуюся непокрытой якорными точками область азаиндольного скаффолда предлагается покрыть классическим фармакофором киназных ингибиторов, подобно тому, как это выполнялось для другого модельного эксперимента, тоже имевшего дело с азаиндольным скаффолдом темплатного лиганда (см. разд. 3.1.3.2, *Ход эксперимента*):

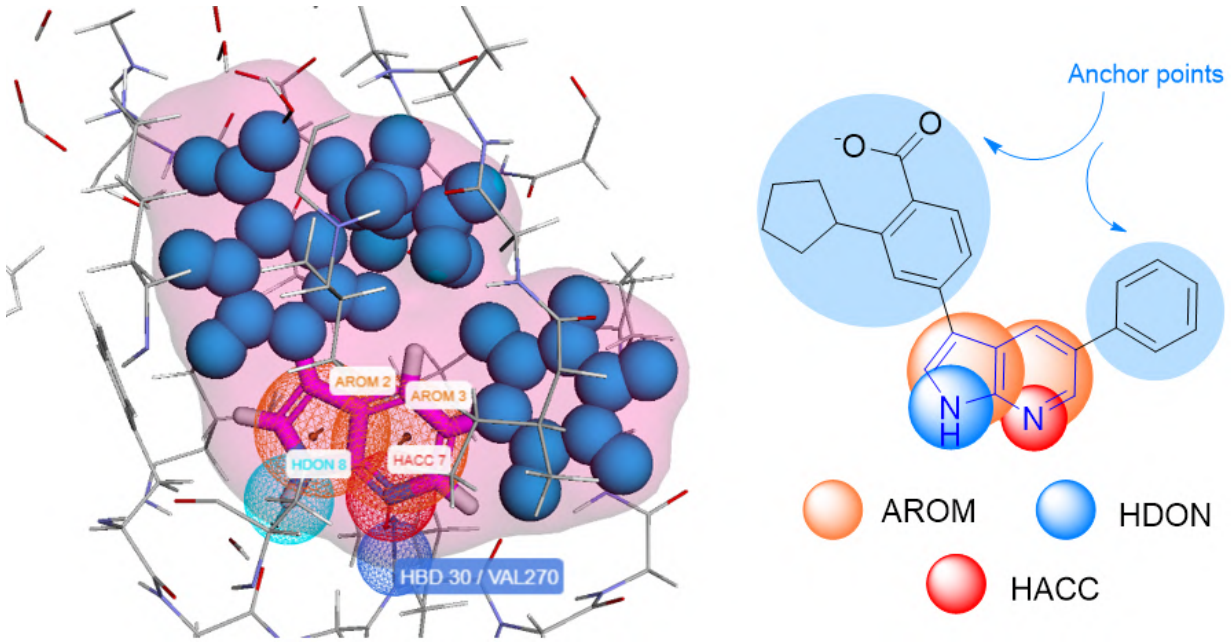


Рисунок 29. Фармакофорная гипотеза и якорные точки для модельного эксперимента №4.

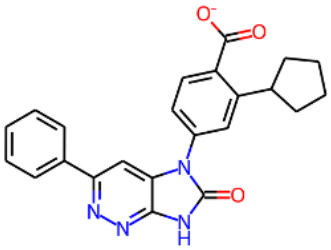
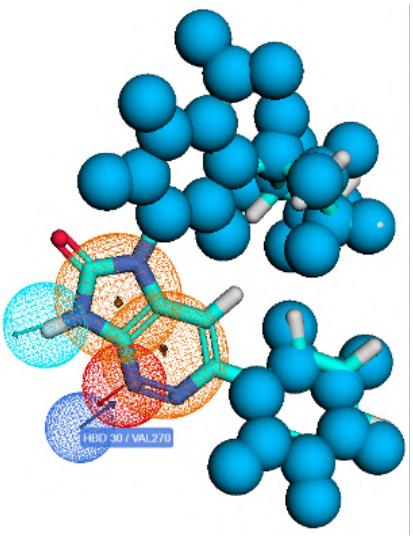
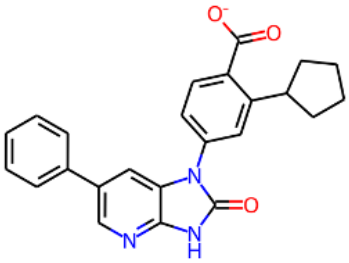
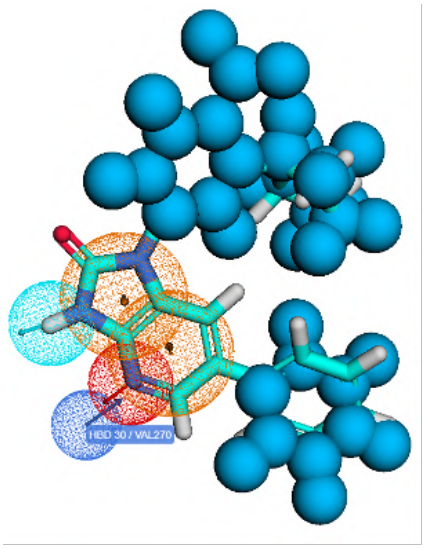
Так же, как и в модельном эксперименте №2 пользователю предлагается указать обязательную точку (донор водородной связи NH-группы аминокислотного остатка в Val270 в *hinge* регионе). В таком случае акцептор со стороны лиганда (HACC) будут хорошо согласованы, что будет способствовать более направленному поиску генеративными моделями наиболее подходящих структур-кандидатов.

Важное замечание перед запуском эксперимента состоит в том, что модуль оценки новизны (Novelty) должен быть отключен, поскольку большая часть молекулярной структуры будет зафиксирована якорными точками, и эта часть будет воспроизводиться во всех генерируемых молекулярных структурах. В противном случае пользователь может получить лишь крайне ограниченные по количеству молекулярных структур результаты.

Эталонные результаты

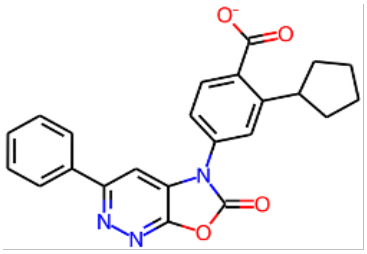
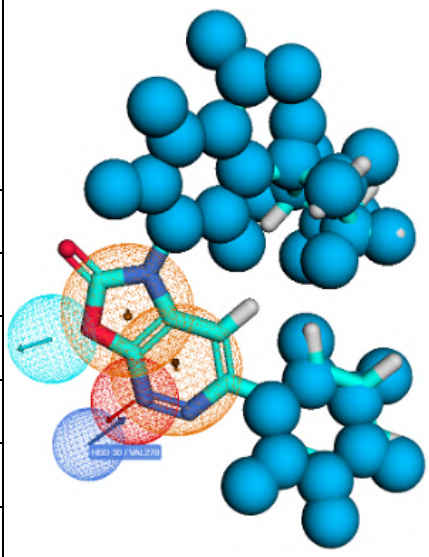
Как и в случае других модельных экспериментов пользователю предлагается в первую очередь ознакомиться с лучшими молекулярными структурами из кластеризованного набора сгенерированных структур (см. табл. 21). Пользователь может убедиться, что закрепленные якорными точками фрагменты сохраняют свою позицию в сгенерированных структурах потенциальных ингибиторов CAMKK2 (Anchor points score), а фармакофорная гипотеза полностью удовлетворяется (Ph4 Score и Ph4 points matched). Более того, для большей части примеров с высокими значениями функции награды, значение оценки синтезируемости ReRSA находится в благоприятном диапазоне 2–3.

Таблица 21. Результаты модельного эксперимента №4

3 лучшие структуры по функции награды (Reward)		
ID	**Структура**	**Наложение на якорные точки**
X1975-8261-0001		
PLI Score	-11.09	
Ph4 Score	0.98	
Ph4 points matched	4/4	
Anchor points score	0.79	
Shape score	0.94	
ReRSA	2.43	
MCE-18	61.29	
Reward	3.38	
X1975-8261-0002		
PLI Score	-11.01	
Ph4 Score	0.99	
Ph4 points matched	4/4	
Anchor points score	0.64	
Shape score	0.92	
ReRSA	2.26	
MCE-18	60.97	
Reward	3.15	

Продолжение на следующей странице

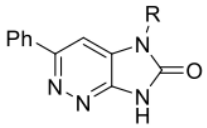
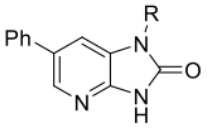
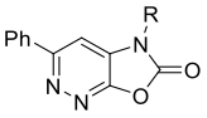
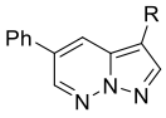
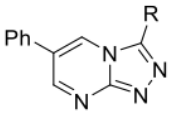
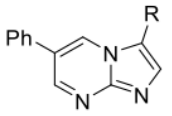
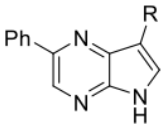
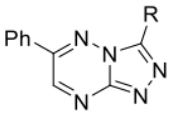
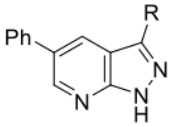
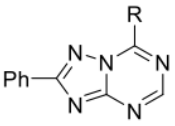
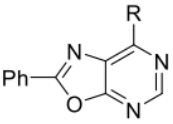
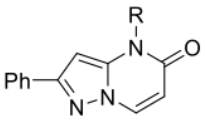
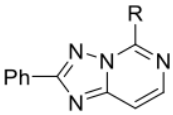
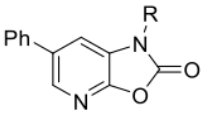
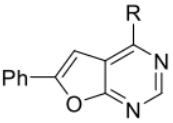
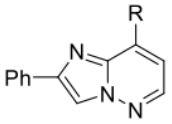
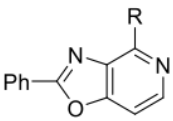
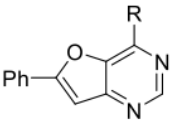
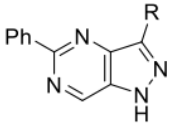
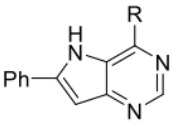
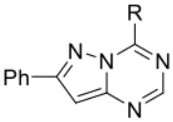
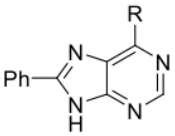
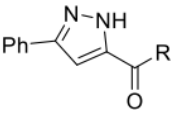
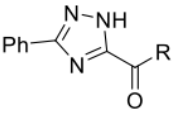
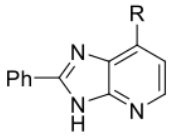
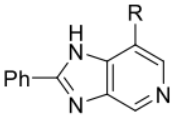
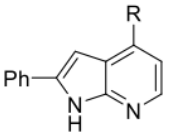
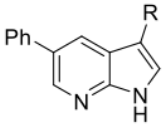
Продолжение таблицы 21

ID	Структура	Наложение на якорные точки
X1975-8261-0003		
PLI Score	-10.96	
Ph4 Score	0.74	
Ph4 points matched	3/4	
Anchor points score	0.78	
Shape score	0.94	
ReRSA	2.43	
MCE-18	61.29	
Reward	3.12	

Тем не менее, как было обозначено в преамбуле, основной целью модельного эксперимента №4 является предоставление как можно большего числа новых скаффолдов, которые ранее не были описаны для ингибиторов CAMKK2 в рамках упомянутого исследования [137].

В общей сложности по результатам модельного эксперимента №4 нам удалось обнаружить 111 уникальных скаффолдов, исключая удлиненные аналоги (см. рис. 30). Ниже мы приводим 28 отфильтрованных скаффолдов (табл. 22), которые были найдены в сгенерированных структурах, и их идентификаторы после применения фильтрации по Anchor Points Score (пороговое значение 0.45) и Ph4 Score (пороговое значение 0.6):

Таблица 22. Отфильтрованные скаффолды в модельном эксперименте №4

 X1975-8261-0001	 X1975-8261-0002	 X1975-8261-0003	 X1975-8261-0005
 X1975-8261-0006	 X1975-8261-0007	 X1975-8261-0008	 X1975-8261-0009
 X1975-8261-0012	 X1975-8261-0013	 X1975-8261-0015	 X1975-8261-0016
 X1975-8261-0017	 X1975-8261-0018	 X1975-8261-0019	 X1975-8261-0021
 X1975-8261-0024	 X1975-8261-0026	 X1975-8261-0028	 X1975-8261-0032
 X1975-8261-0033	 X1975-8261-0043	 X1975-8261-0050	 X1975-8261-0055
 X1975-8261-0063	 X1975-8261-0087	 X1975-8261-0096	 X1975-8261-0110

Нежелательное удлинение скаффолда (см. рис. 30) может возникнуть из-за более высоких порогов максимально допустимого отклонения якорных точек. Если пользователь хочет видеть меньше структур с удлиненным линкером между скаффолдом и якорными точками, то мы рекомендуем ему применить более строгий порог максимально допустимого отклонения якорных точек.

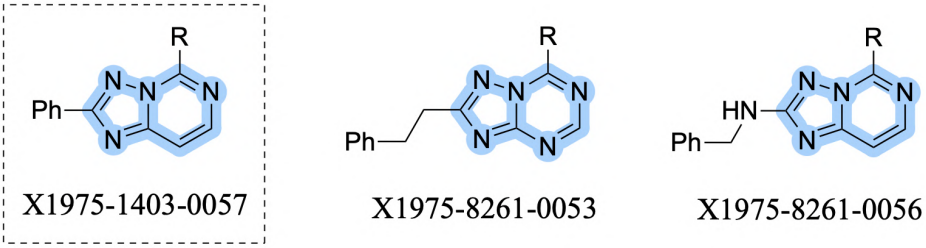
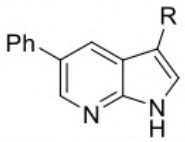
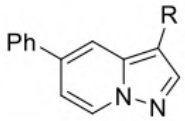
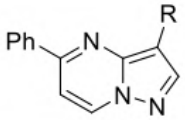
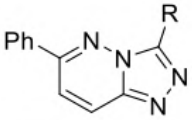
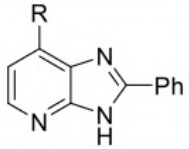


Рисунок 30. Примеры удлиненных скаффолдов (справа). Для сравнения не удлиненный скаффолд — слева.

Как можно убедиться, среди уникальных сгенерированных и отфильтрованных молекулярных структур с уникальными скаффолдами есть те, которые совпадают с теми, что опубликованы в референсном исследовании. Всего в рамках этого запуска модельного эксперимента было найдено 5 таких одинаковых структур/скаффолдов (см. табл. 23).

Таблица 23. Сгенерированные скаффолды, ранее опубликованные в оригинальной статье [137]

ID в статье	Скаффолд	ID в генерации
GSK650394		X1975-8261-0110
19		X1975-8261-0075
20		X1975-8261-0058
25		X1975-8261-0077
34		X1975-8261-0063

Наконец, мы хотели бы обратить внимание на запрет замещений в каждом фрагменте, закрепленном якорными точками (о-циклопентилбензойная кислота и фенил). Так, все атомы водорода были добавлены к якорным точкам. Если же пользователь захочет увидеть какие-либо замещения в этих фрагментах в будущих экспериментах, то мы рекомендуем ему не указывать явные водороды во время настройки якорных точек в положениях, которые он хотел

бы оставить замещенными. Однако в этом конкретном запуске модельного эксперимента №4 мы хотели показать способность платформы выполнять очень сфокусированный и полностью контролируемый *in silico* эксперимент, используя только *scaffold hopping* сценарий моделирования, что может быть полезно при одношаговых исследованиях ССА.

3.1.3.5 Генеративный дизайн заместителей ингибитора MPS1 киназы

Преамбула эксперимента

Известно, что монополярная веретенообразная киназа 1 (MPS1; TTK) является перспективной лекарственной мишенью, регулирующей клеточный цикл. Биологическая роль MPS1 заключается в обеспечении надлежащего прикрепления хромосом к веретену деления во время митоза. Ингибирование активности MPS1 вызывает гибель клеток из-за преждевременного выхода из митоза. Известные ингибиторы MPS1, которые в настоящее время проходят клинические испытания для лечения трижды негативного рака молочной железы, включают NMS-P153, BOS-172722, CFI-402257, BAY1161909 и BAY1217389.

Недавняя статья рассказывает о кампании по оптимизации ингибитора MPS1 **MPI-0479605** [139]. В исследовании описывается поиск заместителей для структуры соединения-лидера при почти полном сохранении скаффолда, представляющего собой кольцо 9H-пурина, соединенное с фенильным кольцом (см. рис. 31). В конечном итоге авторы нашли лучшую комбинацию R-групп (с точки зрения активности), которая наблюдалось для **Соединения 9**.

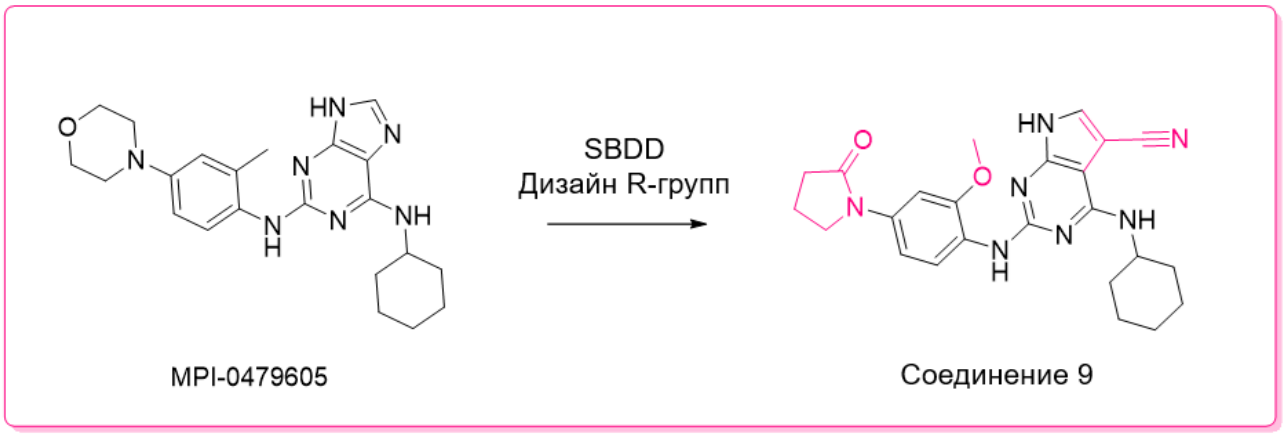
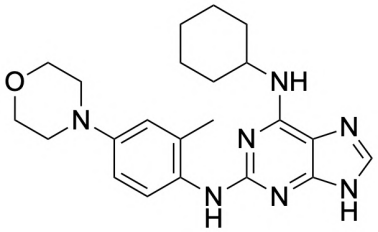


Рисунок 31. Схема оптимизации ингибитора MPS1 [139].

Поскольку количество предложенных комбинаций заместителей ограничено, интерес представляет провести генеративный эксперимент на платформе Chemistry42 с целью обнаружить как можно больше новых вариантов R-групп в рамках той же задачи по

расширению химического пространства вокруг соединения-лидера **MPI-0479605**. Входные данные и ключевые 3D модули для генеративного эксперимента представлены в таблице 24.

Таблица 24. Входные данные и ключевые 3D модули для модельного эксперимента №5

Исходный PDB файл	**Темплатный Лиганд**
5N7V [140]	 **MPI-0479605**
Ключевые 3D модули	
PLI Score Якорные точки Фармакофорный модуль Модуль оценки подобия формы	

Ход эксперимента

Как и в случае модельного эксперимента №4, подлежать закреплению в трёхмерном пространстве будет большая часть молекулярной структуры темплатного лиганда. Однако в этом эксперименте фактически эта область совпадает с областью скаффолда, поэтому нет необходимости добавлять фармакофорные точки, соответствующие ключевым взаимодействиям в области hinge региона. Напротив, фармакофорный модуль можно настроить на генерацию заместителей, если какие-то из уже имеющихся заместителей являются циклическими структурами. Поскольку трансформация **MPI-0479605** в **Соединение 9** проходила с удержанием циклогексанового фрагмента и заменой морфолинового мотива на пирролидоновый, то обе подструктуры (циклогесан и мофролин) можно использовать для создания циклических фармакофорных точек (RING, см. рис. 32).

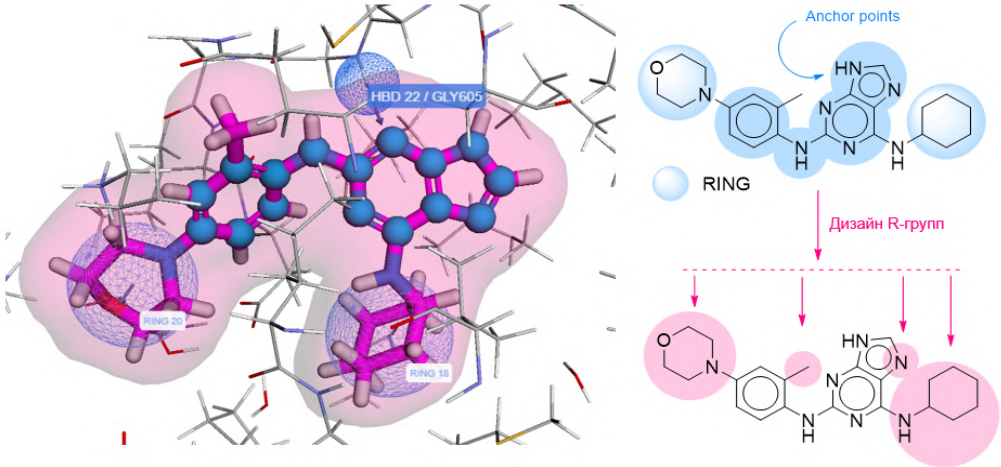


Рисунок 32. Фармакофорная гипотеза и якорные точки для модельного эксперимента №5. Справа розовым цветом подсвечены фрагменты, подлежащие трансформации.

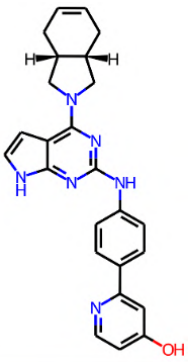
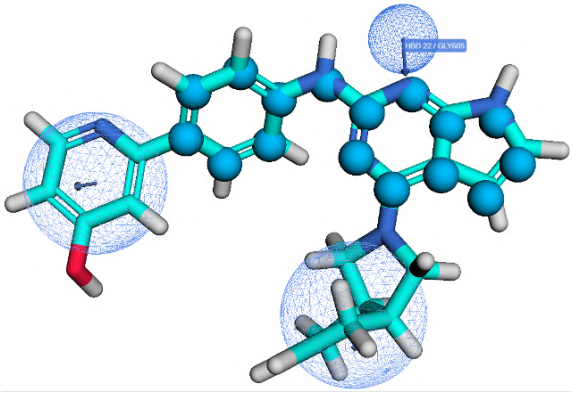
Дизайн соединения **Соединения 9** из **MPI-0479605** предусматривает незначительную трансформацию пуринового скаффолда в 7H-пирроло-[2,3-d]пиримидин, сопровождающейся заменой атома азота на углерод с последующим замещением нитрильной группой. Поскольку целью нашего модельного эксперимента является в том числе и предоставление всех возможностей по структурному моделированию, которые уже ранее были апробированы, то такая замена атома азота на углерод тоже может быть предусмотрена. Как было ранее отмечено (см. разд. 3.1.2) модуль якорных точек позволяет описывать каждую точку при помощи разнообразных атомных примитивов. Таким образом, в описании якорной точки, которая в ходе генеративного эксперимента может претерпеть замену азота на углерод, можно задать два атомных примитива: “ароматический углерод” и “ароматический азот”.

Так же, как и в модельных экспериментах №2 и №4 пользователю предлагается указать обязательную точку — в данном случае, донор водородной связи NH-группы аминокислотного остатка в GLY605 в *hinge* регионе киназы. И поскольку, так же, как и в модельном эксперименте №4 большая часть молекулы является зафиксированной якорными точками, во избежание противоречий с метрикой новизны (Novelty), соответствующий модуль должен быть отключен.

Эталонные результаты

В лучшей, по агрегированной оценке, молекулярной структуре **X1975-8230-0001** можно наблюдать (см. табл. 25) высокие значения оценок модуля якорных точек (качество наложения на якорный фрагмент), благоприятное значение оценки синтезируемости ReRSA и высокое значение MCE-18, очевидно, благодаря замене циклогексанового фрагмента на 2,3,3a,4,7,7a-гексагидро-1H-изоиндольный. Отметим, что уже в этой молекулярной структуре произошла обеспеченная функционалом атомных примитивов замена “ароматического азота” на “ароматический углерод” в области исходного пуринового скаффолда.

Таблица 25. Результаты модельного эксперимента №5

Лучшая структура по функции награды (Reward)		
	Структура	**Наложение на якорные точки**
X1975-8230-0001		
PLI Score	-11.21	
Ph4 Score	0.94	
Ph4 points matched	2/2	
Anchor points score	0.94	
Shape score	0.90	
ReRSA	2.67	
MCE-18	98.23	
Reward	3.50	

Поскольку целью данного исследования является демонстрация возможностей платформы генерировать множество разнообразных R-групп, мы провели подробный анализ R-групп, наблюдаемых в сгенерированном выводе. Мы смогли найти 611 различных заместителей для позиции R1, 7 для R2, 3 для R3, 411 для R4, только один (метил) для R5, 2 для R6, 4 для R7 (см. рис. 33).

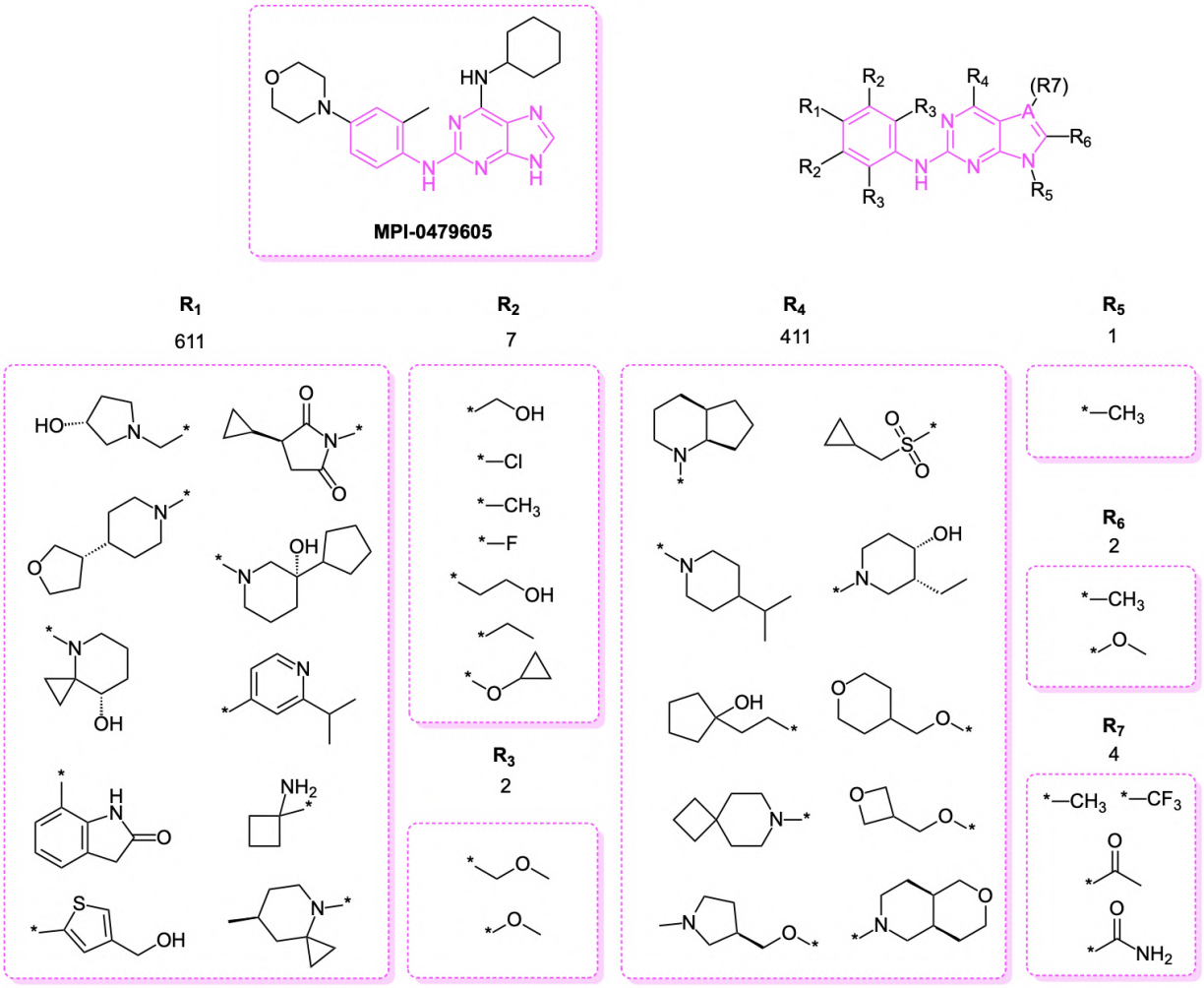
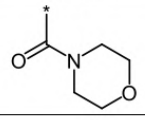
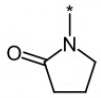
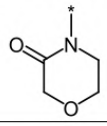


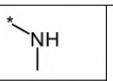
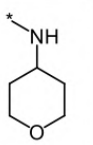
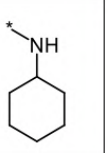
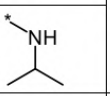
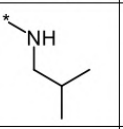
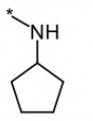
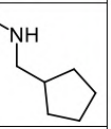
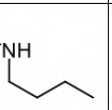
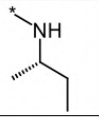
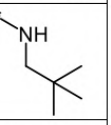
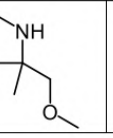
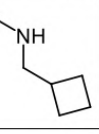
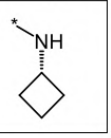
Рисунок 33. Анализ сгенерированных R-групп в модельном эксперименте №5.

Если ставить целью сравнить полученные результаты, с теми, что описаны в референсном исследовании, то можно обнаружить, что меньшая часть из них была предложена платформой (см. рис. 34). Таким образом, как и в случае модельного эксперимента №4, платформа генеративной химии Chemistry42 позволяет изучать ранее неисследованное химическое пространство.

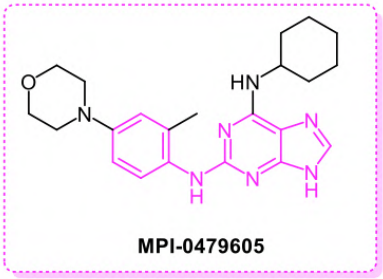
R_1	Сгенерирован
	Нет
	Да
	Да

R_3	Сгенерирован
	Да
	Нет
	Нет

R_7	Сгенерирован
*—CN	Нет
*—CF_3	Да
*—Cl	Нет

R_4		
		
		
		
		
		

R_4: Ни один из заместителей не был сгенерирован.



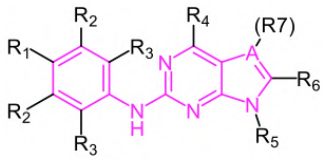


Рисунок 34. Сопоставление сгенерированных заместителей, с предложенными в оригинальной статье [139].

Стоит, однако, обратить внимание на то, что заместители, изучаемые в статье структурно менее разнообразные, чем те, что предлагаются платформой Chemistry42, что, скорее всего, исходя из практики ведения подобных проектов, можно трактовать с позиции ограниченных операционных и синтетических ресурсов.

3.1.3.6 Генеративный дизайн ингибиторов главной протеазы коронавируса SARS-CoV-2 на основе знаний о связывании малого фрагмента

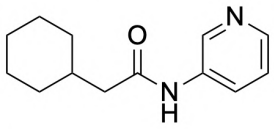
Преамбула эксперимента

Пандемия SARS-CoV-2 в 2020 году явно продемонстрировала неготовность фармацевтической промышленности к острой потребности в эффективных противовирусных препаратах против COVID-19. Тем не менее, пандемия дала толчок беспрецедентным по скорости достижения научным результатам и новые быстрые подходы к ранней разработке лекарств на основе малых противовирусных молекул, связанных с SARS-CoV-2. Одна из самых выдающихся инициатив была предложена Diamond Light Source, которая выпустила сотни сокристаллизованных небольших фрагментоподобных лигандов с основными белками SARS-CoV-2, включая макродомен NSP3 (АДФ-рибозилгидролаза) [141], NSP13 (хеликаза)

[142] и главную протеазу [143], которая является ключевым белком для разработки противовирусных препаратов от COVID-19.

В настоящем модельном эксперименте пользователю предлагается провести *in silico* эксперимент по генеративному дизайну потенциальных ингибиторов главной протеазы коронавируса на основе знаний о связывании малого фрагмента (FBDD). В качестве структурной информации (см. табл. 26) предлагается использовать структуру главной протеазы в комплексе с фрагментоподобным лигандом **Z31792168**, полученную в на синхротроне Diamond Light Source [143].

Таблица 26. Входные данные и ключевые 3D модули для модельного эксперимента №6

Исходный PDB файл	**Темплатный Лиганд**
5R84 [144]	 **Z31792168**
Ключевые 3D модули	
PLI Score Якорные точки Фармакофорный модуль	

Ход эксперимента

Особенности конфигурации генеративного эксперимента на платформе Chemistry42 для реализации FBDD-сценария заключаются в том, что весь фрагмент или его часть предстоит закрепить якорными точками. Но поскольку размер генерируемых молекулярных структур будет значительно превышать размер исходного фрагмента, то следует учесть два важных момента:

1. Область допустимого для генерации трёхмерного пространства должна быть значительно увеличена (по умолчанию генерация допускается в область ограниченную пространственно лигандом);
2. Модуль подобия формы должен быть отключен. В противном случае награждаться будут только те структуры, которые напоминают по форме темплатный фрагментоподобный лиганд.

В остальном логика конфигурации эксперимента передает уже ранее описанные идеи. Пользователю предлагается зафиксировать в трёхмерном пространстве при помощи якорных точек пиридиновый цикл, обеспечивающий важное связывание с остатком гистидина HIS63, а парой фармакофорная точка-обязательная точка акцентировать важную водородную связь между остатком глутаминовой кислоты GLU166 и кислородом амидной группы фрагмента (см. рис. 35). Помимо этого, исходя из анализа подпакета S2 сайта связывания протезы, благоприятную роль в связывании играют гидрофобные фрагменты, представленные в **Z31792168** циклогексановым заместителем. Эта область может быть снабжена гидрофобной фармакофорной точкой (HYD).

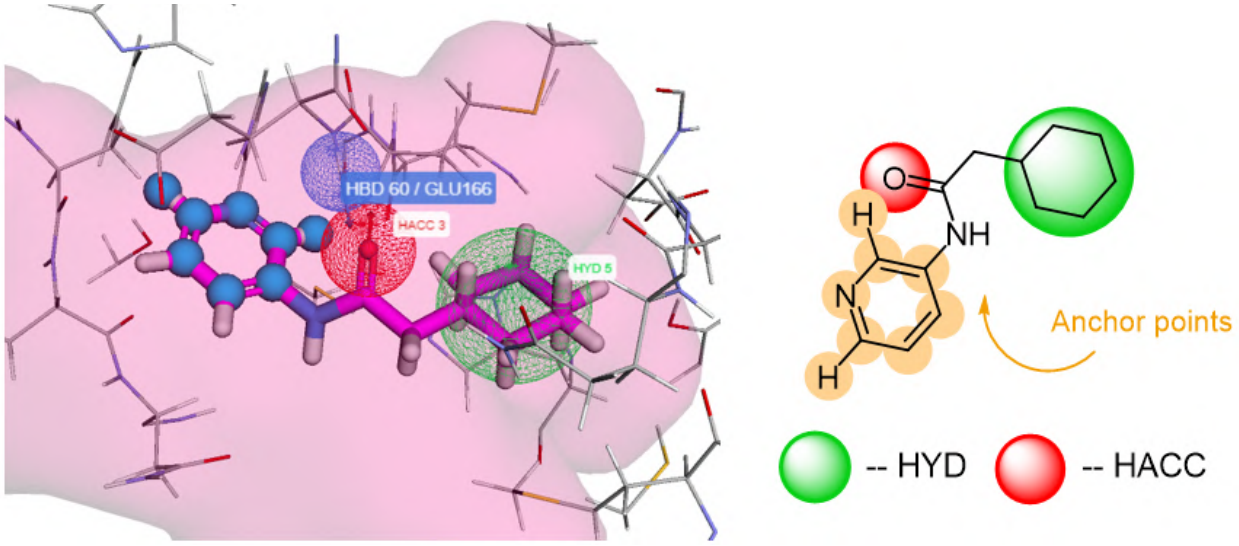
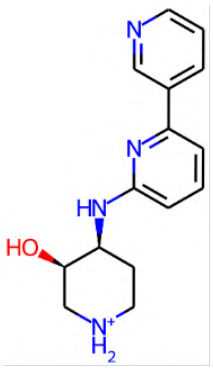
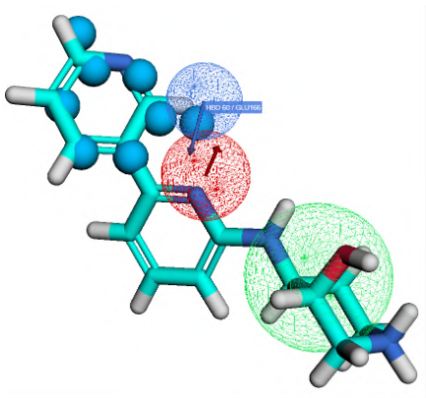


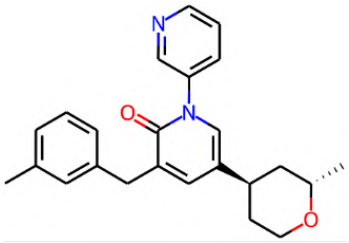
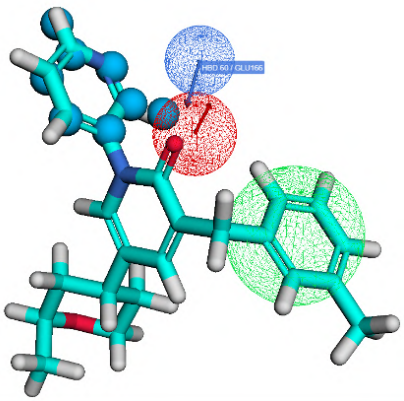
Рисунок 35. Фармакофорная гипотеза и якорные точки для модельного эксперимента №6.

Эталонные результаты

Полученные в результате модельного эксперимента молекулярные структуры демонстрируют способность платформы Chemistry42 создавать релевантное химическое пространство для FBDD-сценария молекулярного моделирования. Благодаря расширенной допустимой области генерации и отключению модуля подобия формы, часть структур значительно превышает по размеру исходный фрагмент **Z31792168**. В случае одной из лучших структур **X1975-4541-2453** (см. табл. 27) удалось получить интересный результат амидного биоизостеризма, в котором исходный амидный мотив трансформировался в пиридоновый с удержанием соответствия фармакофорным точкам (в первую очередь акцептору HACC).

Таблица 27. Результаты модельного эксперимента №6

Лучшая структура по функции награды (Reward)		
	Структура	**Наложение на якорные точки**
X1975-4541-0002		
PLI Score	-8.37	
Ph4 Score	0.83	
Ph4 points count	2/2	
Anchor points score	0.92	
ReRSA	2.19	
MCE-18	49.00	
Reward	3.40	

Лучшая структура по PLI Score		
	Структура	**Наложение на якорные точки**
X1975-4541-2453		
PLI Score	-9.95	
Ph4 Score	0.92	
Ph4 points count	2/2	
Anchor points score	0.95	
ReRSA	4.00	
MCE-18	69.56	
Reward	2.77	

В настоящем модельном эксперименте мы предлагаем пользователю особое внимание обратить на доступное разнообразие модификаций пиридинового цикла, закрепленного якорными точками (см. рис. 36). Обилие как одинарных замещений, так и сложных, разветвленных, а также конденсированных циклов демонстрирует возможности платформы генеративной химии эффективно исследовать химическое пространство вокруг фрагментоподобных лигандов.

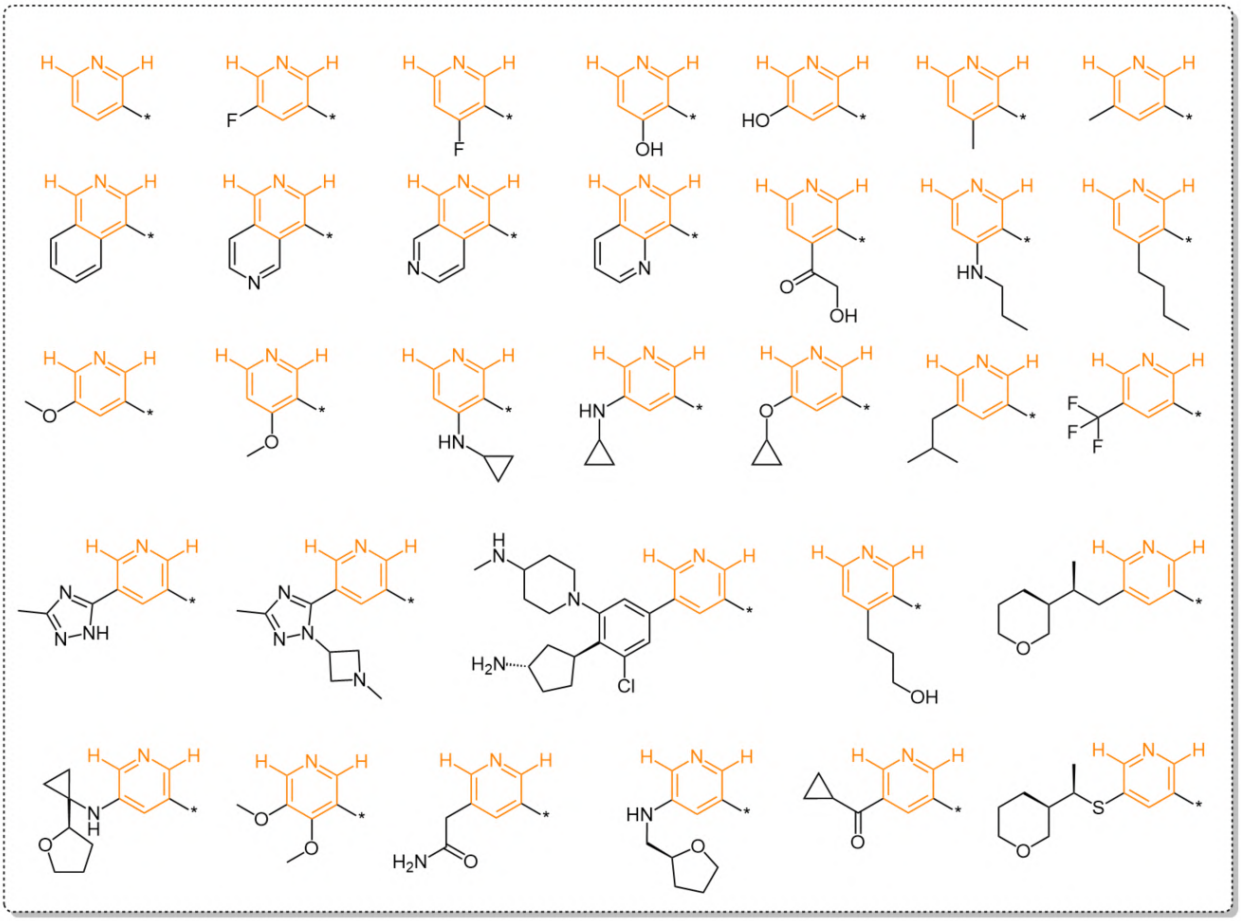


Рисунок 36. Разнообразие замещений по пиридиновому фрагменту, закрепленному с помощью якорных точек в модельном эксперименте №6.

3.1.3.7 Дальнейшее развитие практики модельных экспериментов в рамках платформы Chemistry42

Несмотря на доступное обилие сценариев по молекулярному моделированию потенциальных лекарственных веществ, которые доступны на платформе, как показывает практика модельных экспериментов, не все сценарии (см. разд. 1.1.2) могут быть покрыты на данный момент.

Серьезным ограничением на пути к реализации сценария дизайна линкера между двумя несвязанными фрагментами является техническая невозможность задать информацию

одновременно о двух темплатных лиганда (фрагментоподобных лигандах), поскольку конфигурация платформы ограничивается только одним таким лигандом. По тем же соображениям не может быть проведен генеративный эксперимент по моделированию протеолитических химерных соединений (PROTAC).

Сценарий макроциклизации на данный момент также не может быть эффективно использован на платформе из-за несовершенства модуля генерации конформаций, который не несет в себе знания о конформационном пространстве макроциклов.

Тем не менее, вышеперечисленные технические ограничения не являются непреодолимыми, и в будущем планируется провести необходимые изменения, позволяющие реализовывать новые базовые сценарии молекулярного моделирования и оформить их реализацию в виде новых отдельных модельных экспериментов.

3.1.4 Реальные примеры использования платформы Chemistry42 для практического решения задач медицинской химии

Если изучить разделы о модельных экспериментах, демонстрирующих функционал платформы генеративной химии, то может сложиться ложное впечатление, что результаты работы платформы не подтверждены экспериментально в рамках биологических тестирований и клинических испытаний, что зачастую наличествует в публикациях о новых моделях генеративной химии и что нашло отражение в критическом обзоре нашей научной группы, посвященном проблемам современной генеративной химии [21]. Однако, этому противоречит обширная практика использования инструментов платформы Chemistry42, как многочисленными клиентами компании, так и командами разработки потенциальных лекарственных веществ группы компаний Insilico Medicine. Но поскольку разработки клиентов, так же, как и некоторые внутренние разработки, носят строго коммерческий и закрытый характер, упоминанию подлежат те примеры, которые удостоились публикации.

Активными пользователями платформы является команда коллег из шанхайского отделения Insilico Medicine. Ими за последние годы был опубликован целый ряд статей, иллюстрирующих то, как при помощи функционала Chemistry42 могут быть проведены кампании по разработке потенциальных лекарственных веществ. В рамках этих проектов разрабатывались ингибиторы CDK8, SIK2, Polθ, PHD2, CDK18 (см. рис. 37)

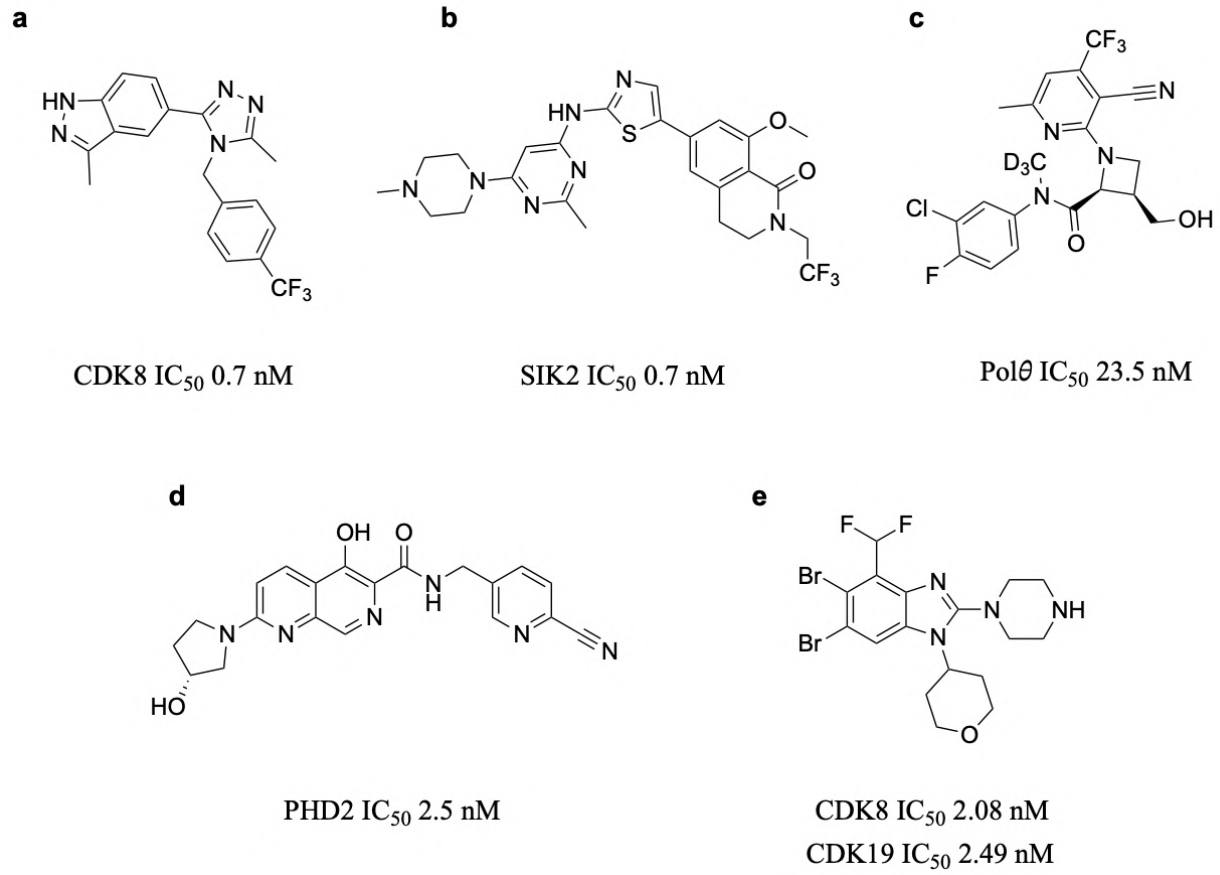


Рисунок 37. Низкомолекулярные ингибиторы, разработанные с помощью платформы Chemistry42. (**a)** Ингибитор циклин-зависимой киназы 8 (CDK8) [145], **(b)** Ингибитор соль-индуцируемой киназы 2 (SIK2) [146], **(c)** Ингибитор ДНК-полимеразы низкой точности тета (Polθ) [147], **(d)** Ингибитор фермента домена пролилгидроксилазы 2 (PHD2) [148], **(e)** Ингибитор циклин-зависимых киназ 8 и 10 (CDK8/10) [149].

Подробнее остановимся на нескольких кампаниях разработки потенциальных лекарственных молекул с применением платформы Chemistry42.

3.1.4.1 Идентификация соединения-хита в ходе дизайна ингибиторов CDK20

Впервые при помощи платформы было продемонстрировано, что алгоритмы генеративной химии сочетаемы с успехами в области прогнозирования фолдинга белков, у которых нет информации о трехмерных структурах, кроме как апо-структура, спрогнозированная при помощи алгоритма AlphaFold2 [150]. В качестве объекта исследования была взята циклин-зависимая киназа 20 (CDK20), для которой в PDB нет доступных структурных данных. Спрогнозированную алгоритмом AlphaFold2 [8] апо-структуру CDK20 использовали для *de novo* дизайна на основе структуры мишени (см рис. 38).

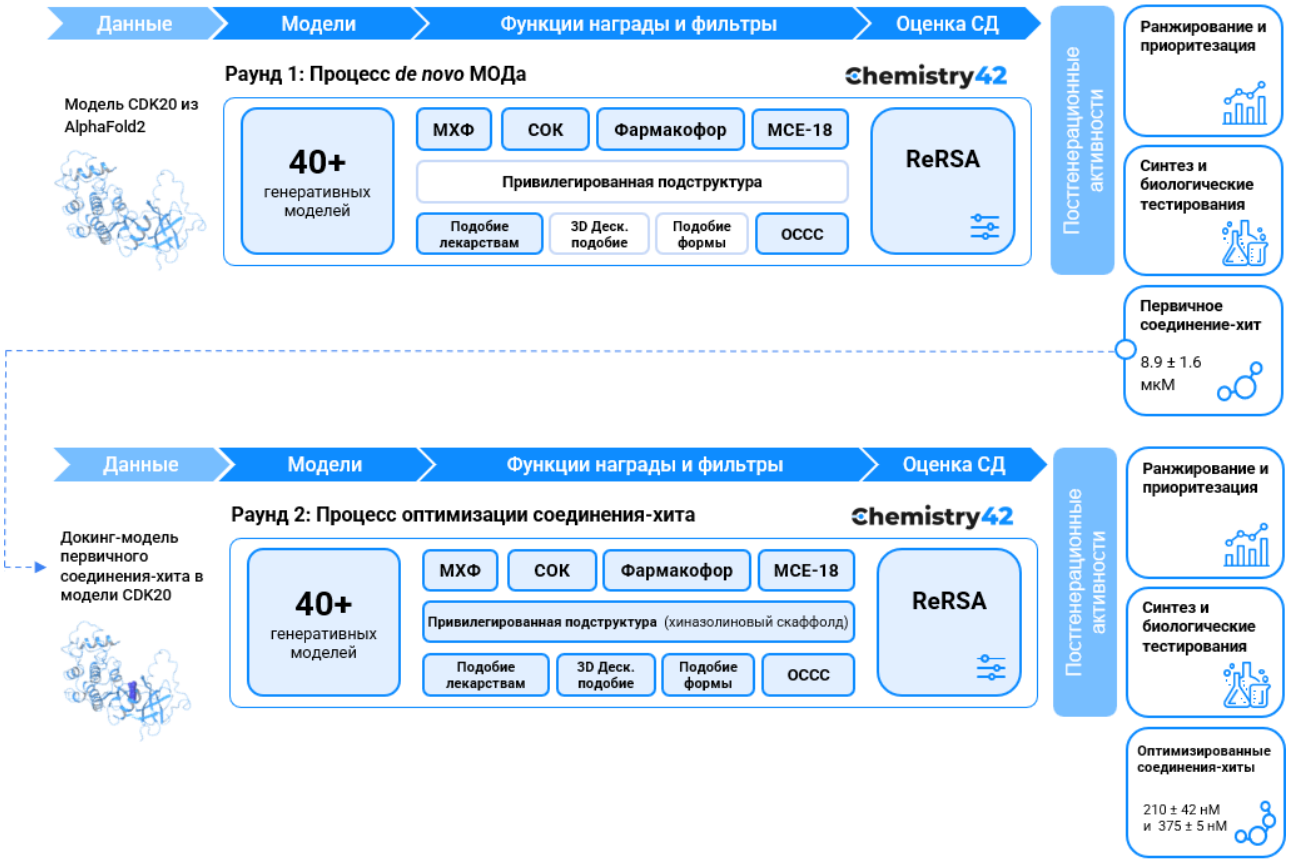
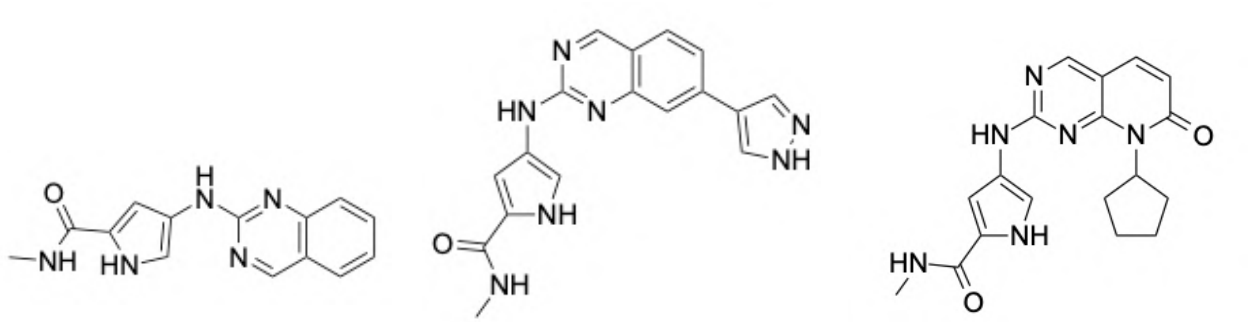


Рисунок 38. Схема генеративного дизайна ингибиторов CDK20 на основе модели AlphaFold2.

Помимо этого, была смоделирована простейшая двухточечная фармакофорная гипотеза, соответствующая киназным ингибиторам. Вспомогательную роль в направлении генерации в нужное физико-химическое пространство играл модуль СОК, сфокусированный на химическом пространстве ингибиторов киназ из семейства CDK. Результаты первого раунда генеративного дизайна были кластеризованы на платформе, а лучшие представители кластеров отправлены на синтез. По итогам синтетической кампании большую часть молекул удалось синтезировать, после чего они были отправлены на исследование *in vitro* ингибирования CDK20. Полученное соединение-хит **ISM042-2-001** со значением активности 9.2 мкМ (K_d) (см. рис. 39) было затем использовано в оптимизационном генеративном эксперименте, подобном тому, что моделировался при помощи модельного эксперимента №3 (см. разд. 3.1.3.3), в ходе которого фиксировалась важный для связывания фрагмент соединения-хита при помощи якорных точек. По итогам оптимизационной кампании удалось получить два оптимизированных соединения с улучшенными показателями активности K_d: **ISM042-2-048** (567 нМ) и **ISM042-2-049** (360 нМ), структуры которых приведены на рис. 39.

А **Б** **В**



ISM042-2-001
CDK20 K_d 9.2 мкМ

ISM042-2-048
CDK20 K_d 567 нМ

ISM042-2-049
CDK20 K_d 360 нМ

Рисунок 39. А. Соединение-хит ингибитор CDK20, сгенерированный платформой Chemitsry42. **Б** и **В** Оптимизированные соединения-хиты.

Результаты генеративного эксперимента были критически оценены нашей научной группой с точки зрения совместимости моделей из AlphaFold2 с движками генеративной химии [9].

3.1.4.2 Ранняя разработка ингибитора главной протеазы коронавируса SARS-CoV-2

Автор настоящей диссертационной работы непосредственно руководил ранней разработкой малых лекарственных в молекул в рамках проекта по созданию вышеупомянутого ингибитора **ISM-3312** главной протеазы коронавируса SARS-CoV-2 при помощи ранней версии платформы Chemistry42, проводил все эксперименты по молекулярному моделированию, отбор структур-кандидатов на синтез и анализ биологических данных с обратной связью в процесс моделирования. По этой причине данный проект представляет особый интерес с точки зрения демонстрации того, как со-разработчик программного обеспечения может впоследствии работать с этим программным обеспечением как конечный пользователь.

В самом начале пандемии существовало очень мало исходных данных для использования в рамках SBDD-парадигмы дизайна ингибиторов M^{pro} SARS-CoV-2. Первая опубликованная со-кристаллическая структура M^{pro} SARS-CoV-2 (PDB: 6LU7) была представлена 5 февраля 2020 года с ковалентным ингибитором N3 [151] (рис. 40). Затем 25 марта 2020 года на сайте PDB была выложена структура 6W63 [152], содержащая нековалентный ингибитор X77 (рис. 40).

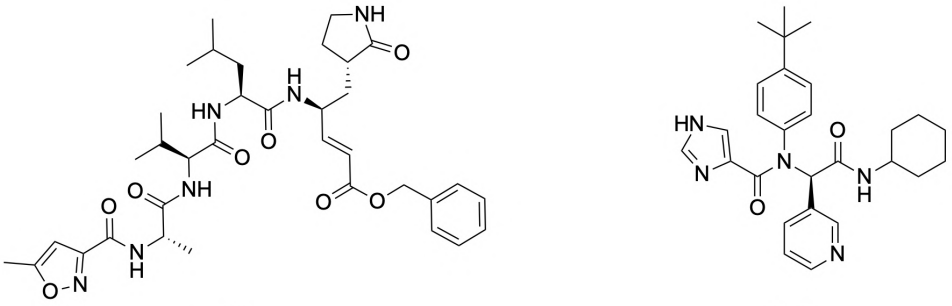


Рисунок 40. Первые описанные в виде ко-кристаллов лиганды SARS-CoV-2 M^{pro}.

Детальный анализ хемотипа соединения X77, проведённый в предыдущих работах [153,154] группы А. Месекара, посвященных ингибиторам главных протеаз SARS-CoV и MERS, позволил нам понять значительный потенциал данного хемотипа, который заключается в исключительно высокой синтетической доступности скаффолда, образованного при помощи многокомпонентной реакции Уги, что обеспечивает быструю разработку хемотипа и его оптимизацию от стадии соединения-хита до соединения-лидера.

Для обнаружения соединения-хита на платформе Chemistry42 был проведен виртуальный скрининг подмножества соединений библиотеки компании ChemDiv [112], представляющих продукты реакции Уги. В конфигурации виртуального скрининга были учтены следующие ключевые для эффективного связывания взаимодействия в виде фармакофорных точек на основе лиганда из со-кристалла 6W63 (см. рис. 41А):

1. Акцептор водородной связи для с остовом Glu166;
2. Ароматическая точка для обеспечения потенциального π-стэкинга с His41 в отобранных лигандах. Та же точка была задана с логикой «ИЛИ» и может также вести себя как гидрофобная точка благодаря близкому расположению гидрофобного S2-подкармана;
3. Гидрофобная точка, занимающая гидрофобный S2-подкарман протеазы;
4. Гидрофобная точка для занятия гидрофобного S4-подкармана протеазы;

и **обязательная точка** (см. разд. 3.1.2) в виде донора водородной связи на имидазоле остатка His163.

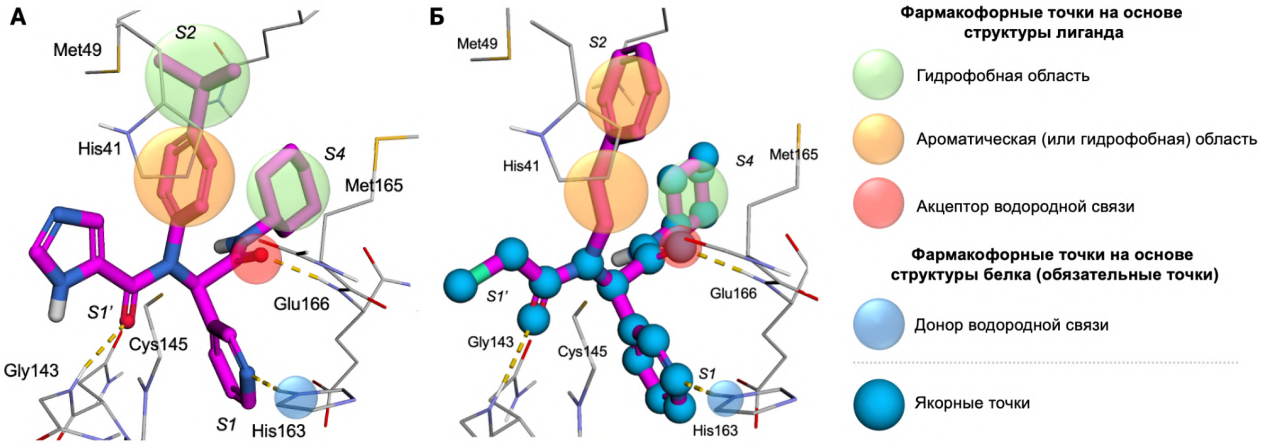


Рисунок 41 А. Конфигурация фармакофорных точек на основе соединения X77 для проведения виртуального скрининга на платформе Chemistry42 в целях обнаружения соединений-хитов, PDB: 6W63. **Б.** Конфигурация фармакофорных и якорных точек на платформе Chemistry42 в рамках генеративного эксперимента по исследованию активного хемотипа в пределах оптимизона, соответствующего аминной компоненте в реакции Уги, на основе докинг-позы соединения-хита **INSCoV-001**.

В ходе виртуального скрининга и последующего *in vitro* скрининга виртуальных соединений-хитов была обнаружена серия из 5 активных ингибиторов M^{pro} SARS-CoV-2 с диапазоном активности IC_{50} 0.98–8.65 мкМ, представляющих α-хлорацетамиды, получаемые по реакции Уги из α-хлоруксусной кислоты. Было высказано предположение, что серия активных α-хлорацетамидов представляет собой ковалентные ингибиторы, поскольку α-хлорацетамиды являются выраженными электрофильными агентами в реакции с остатками цистеина в активных сайтах цистеиновых протеаз. Доказательство ковалентного связывания представителей активного хемотипа предстояло продемонстрировать посредством кристаллографического эксперимента. Наиболее активное соединение-хит **INSCoV-001** со значением активности 0.98 мкМ (IC_{50}) в протеазном эссее было использовано в качестве начальной точки для дальнейшей оптимизации хемотипа.

В ходе наработки (*hit expansion*) и оптимизации химического пространства вокруг соединения-хита была улучшена активность в протеазном эссее и ряд *in vitro* ADME свойств для представителей хемотипа. Моделирование структур на данном этапе осуществлялось посредством генеративных экспериментов в ходе которых использовалась логика оптимизонов [155]. В общей сложности рассматривалось 4 оптимизона, каждый из которых соответствовал отдельному компоненту из 4-компонентной реакции Уги: изонитрилу, альдегиду, амину и кислоте (см. рис. 42). В то же время, в рамках этого этапа разработки,

оптизон кислоты не изменялся, поскольку связывающий фрагмент молекулы (*warhead*) был неотъемлемой частью α-хлоруксусной кислоты.

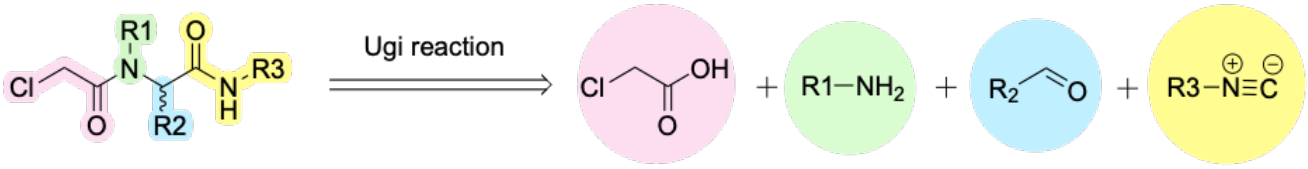
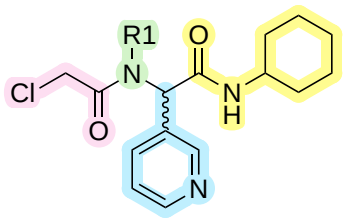
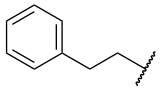
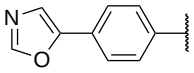
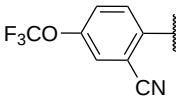
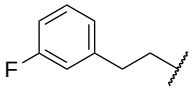


Рисунок 42 Ключевые оптимизоны в рамках компании по наработке и оптимизации химического пространства активных соединений и их синтетические предшественники в реакции Уги: кислотный оптимизон (розовый), аминовый R1-оптимизон (зелёный), альдегидный R2-оптимизон (голубой) и изонитрильный R3-оптимизон (жёлтый).

В то время как один из оптимизонов подвергался исследованию, остальные закреплялись при помощи **якорных точек** (см. разд. 3.1.2). Таким образом удавалось делать точечные изменения в молекулярных структурах в целях формирования консистентных пространств структура-активность/структура-свойство(а). Так, например, в ходе первого генеративного цикла по расширению-оптимизации хемотипа производилось исследование аминного оптимизона, исходя из модели, полученной на основе докинг-позы соединения-хита **INSCoV-001**, в то время как остальная часть молекулы закреплялась при помощи якорных точек (см. рис. 41Б). В ходе данного генеративного цикла удалось установить , что наиболее благоприятными аминными компонентами в рамках соответствующего оптимизона с точки зрения активности в протеазном эссее (M^{pro} SARS-CoV-2, IC_{50}) являются 5-трифторметокси-,2-аминобензонитрил (**INSCoV-517**), 4-(оксазол-5-ил)анилин (**INSCoV-501**) и 2-(3-фторфенил)этан-1-амин (**INSCoV-558**) (см. табл. 28).

Таблица 28. Наиболее активные в протеазном эссее соединения, полученные в ходе первого оптимизационного цикла

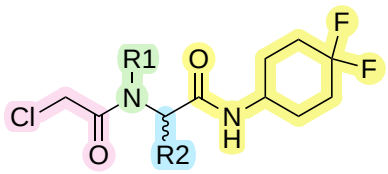


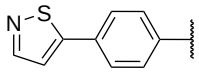
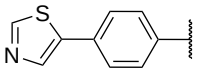
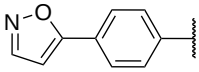
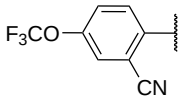
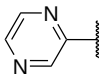
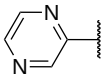
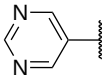
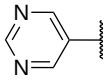
ID	**INSCoV-001**	**INSCoV-501**	**INSCoV-517**	**INSCoV-558**
Свойства \ R1				
M^{pro} SARS-CoV-2, IC_{50} (нМ)	980	680	242	399
ЧПМ/МПМ Cl_{int} (мкМ/мин/мг)	––	205.0/212.8	––	665.8/420.1
Caco-2 проницаемость, (10^{-6} см/с)	––	0.15	––	0.23
Растворимость в воде, pH 7.4 (мкМ)	––	2.88	––	8.51
CYP P450 1A2/2C9/2C19/2D6/3A4, IC_{50} (мкМ)	––	>50/7.59/5.79 /6.75/0.59	––	>50/9.22/7.50 />50/1.49

Последующие генеративные циклы в рамках других оптимизонов позволили значительно улучшить профиль активности в протеазном эссее для хемотипа и основные *in vitro* ADME-параметры (см. табл. 29):

1. Метаболическая стабильность соединения в микросомальной фракции печеночных микросом человека (ЧПМ) и мыши (МПМ), измеренная в терминах клиренса;
2. Проникающая способность соединения через монослой клеток Caco-2;
3. Растворимость соединения в воде;
4. Профиль ингибирования соединением цитохромов P450, ключевых для метаболизма ксенобиотиков;

Таблица 29. Избранные соединения из оптимизационной кампании и их свойства



ID	INSCoV-601I	INSCoV-601G	INSCoV-600J	INSCoV-517A
R1				
R2				
M^{pro} SARS-CoV-2, IC_{50} (нМ)	62	81	106	86
ЧПМ/МПМ Cl_{int} (мкМ/мин/мг)	123.7/629.9	46.2/437.6	17.0/27.8	23.6/22.4
Caco-2 проницаемость, (10^{-6} см/с)	0.04	0.04	0.002	0.31
Растворимость в воде, pH 7.4 (мкМ)	<1.6	<1.6	34.5	5.31
CYP P450 1A2/2C9/2C19/2D6/3A4, IC_{50} (мкМ)	>50/11.3/9.5 /5.5/0.3	>50/20.8/4.6 /3.5/0.6	>50/>50/49.9 /9.4/43.2	>50/>50/>50 />50/>50

Во время разработки характеризовались преимущественно рацематы, однако для тех соединений, активность рацематов которых попадала в диапазон значений IC_{50} менее 100 нМ, проводилось разделение стереоизомеров. Так, для рацемического соединения **INCoV-601I** был получен активный стереоизомер **INCoV-601I-R** со значением IC_{50} 50 нМ и предварительно установленной абсолютной R-конфигурацией стереоцентра по результатам молекулярного моделирования на платформе Chemistry42. Отметим, что R-конфигурация присваивалась всем молекулярным структурам, поскольку таковая была более благоприятна с точки зрения моделирования связывания в активном сайте протеазы. Для подтверждения абсолютной конфигурации стереоцентра в соединении **INCoV-601I-R** и подтверждении ковалентного характера связывания оно было со-кристаллизовано с протеазой, а со-кристалл охарактеризован при помощи кристаллографического метода.

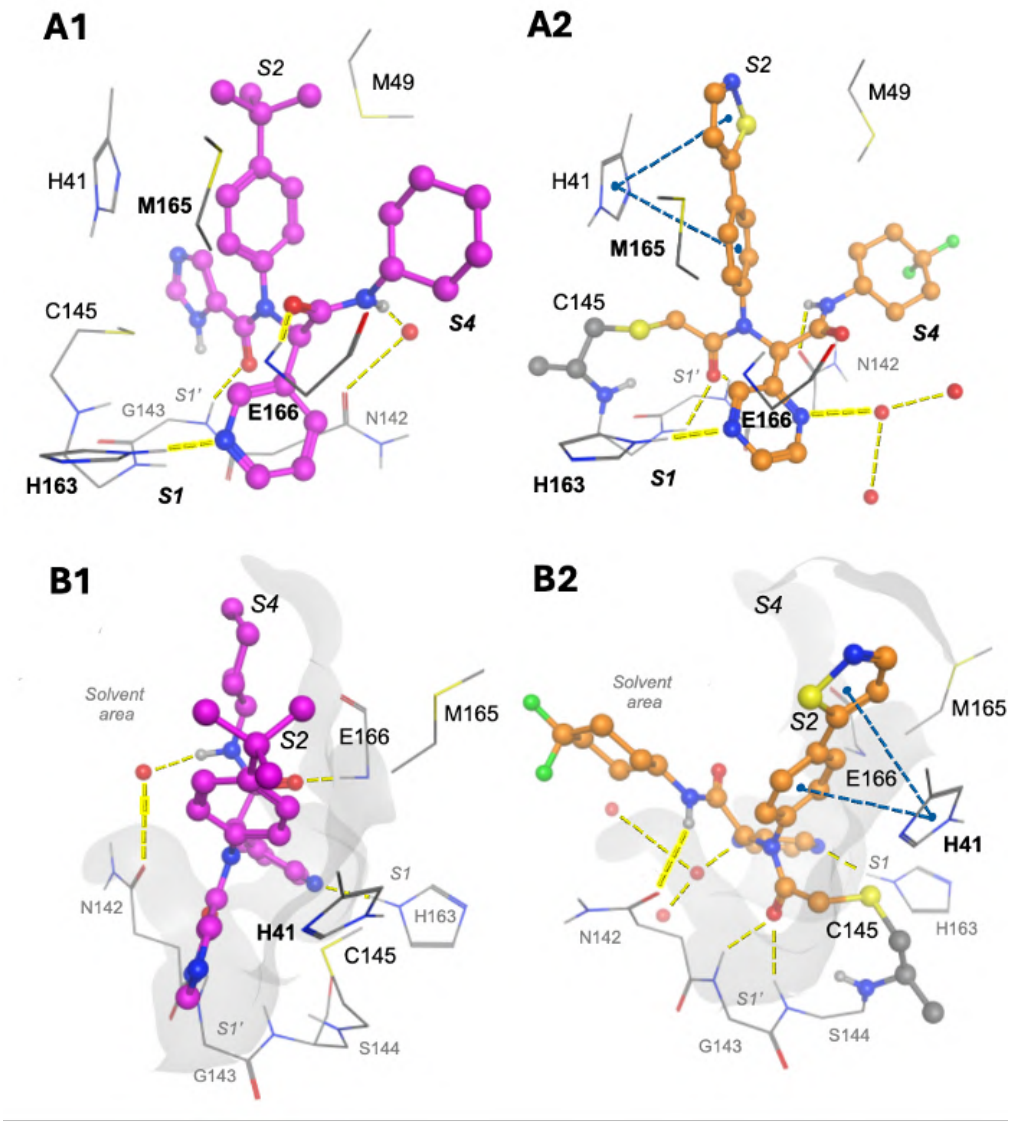


Рисунок 43. Виды **A1** и **B1**. Режим связывания нековалентного соединения **X77** (пурпурный), PDB: 6W63. Виды **A2** и **B2**. Режим связывания ковалентно связанного соединения **INCoV-601I-R** (оранжевый), PDB: 9WHE. Жёлтые пунктирные линии — водородные связи, тёмно-синие пунктирные линии — π-стэкинг.

Рентгеноструктурное исследование подтвердило ковалентное связывание с остатком Cys145 (см. рис. 43 A2 и B2); однако фактический режим связывания соединения **INCoV-601I-R** (PDB: 9WHE) частично отличался от смоделированного, который имитировал режим связывания соединения X77 (см. рис. 43 A1 и B1). Ожидаемое занятие S4-подкармана 4,4-дифторметилциклогексильным заместителем в соединении **INCoV-601I-R**, аналогичное тому, как циклогексильная группа занимает S4 в случае соединения X77, не наблюдается. Вместо этого вся область R3-оптимизона соединения **INCoV-601I-R** экспонируется в область растворителя.

Амидный фрагмент R3-оптимизона соединения **INCoV-601I-R** непосредственно связывается с остатком Asn142 через водородную связь, тогда как у соединения **X77** связывание с Asn142 происходит через водяной мостик, что позволяет ему образовывать водородную связь с остовом Glu166. Прямое связывание с Asn142 у соединения **INCoV-601I-R** также приводит к потере водородной связи с Glu166, тогда как это взаимодействие было выбрано в качестве основы для одной из фармакофорных точек (см. рис. 41) и части выбранных якорных точек.

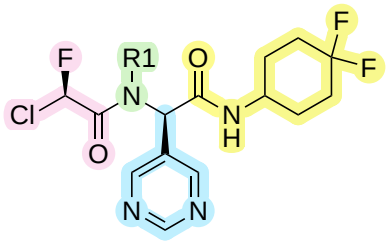
Также нам удалось наблюдать более выраженный, чем в случае соединения **X77**, параллельный π-стэкинг между боковой цепью His41 и биарильным фрагментом соединения **INCoV-601I-R**. Наличие второго атома азота в пиразиновом кольце R2-оптимизона соединения **INCoV-601I-R** позволило ему связываться с цепочкой из молекул воды, что не наблюдается у соединения **X77**. Наконец, ковалентно связанное соединение **INCoV-601I-R** прочно фиксируется в S1’-области сайта связывания протеазы благодаря водородным связям с остовами Gly143 и Ser144, тогда как соединение **INCoV-601I-R** связывается там только остовом остатка Gly143.

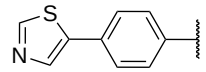
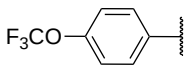
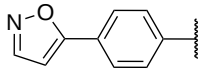
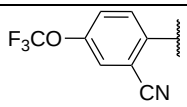
Несмотря на то, что нам удалось оптимизировать ряд свойств, таких как активность в протеазном эссее, метаболическая стабильность в микросомальной фракции печени и профиль ингибирования цитохромов P450, другие свойства, то как, растворимость в воде, проникающая способность не были должным образом улучшены. Более того, одно из свойств, которое обычно не рассматривается в базовом наборе *in vitro* ADME эссеев — стабильность в плазме крови, препятствовало дальнейшему развитию серии в целом. Так, например, период полураспада соединения **INCoV-601I-R** в плазме составлял всего 4.2 мин. В целом, такая нестабильность α-хлорацетамидной серии соединений в нуклеофильных условиях плазмы крови объясняется крайне высокой электрофильностью связывающего фрагмента молекулы (СФМ).

Посредством анализа литературных источников [156] нам удалось найти эффективную альтернативу α-хлорацетамидному СФМ в виде (R)-α-фтор,α-хлорацетамидного фрагмента (ФХА), который может быть лаконичным образом инкорпорирован в хемотип путем замены α-хлоруксусной кислоты на (R)-α-фтор,α-хлоруксусную кислоту. В результате проблема нестабильности соединений в плазме крови была в большинстве случаев нивелирована, при том, что негативного эффекта на активность в протеазном эссее не наблюдалось. Стабильность ФХА-фрагмента в физиологических условиях объяснима с точки зрения SN2-механизма, который делает α-атом углерода СФМ стерически менее доступным для нуклеофилов в плазме

крови. Более того, путем замены СФМ удалось улучшить растворимость соединений, их метаболическую стабильность и проникающую способность. Последнее, по-видимому, объясняется несколько более высокой липофильностью ФХА-фрагмента. Таким образом, заметное улучшение профиля свойств хемотипа посредством внедрения нового СФМ позвонило номинировать серию соединений-лидеров, представленных в табл. 30.

Таблица 30. Избранные представители α-фтор,α-хлорацетамидной лидирующей серии ингибиторов M^{pro} SARS-CoV-2



ID	INSCoV-614-R	INSCoV-517D-R	INSCoV-614A-R	INSCoV-517C-R
R1				
M^{pro} SARS-CoV-2, IC_{50} (нМ)	61±1	121±7	128±3	69±3
ЧПМ/МПМ Cl_{int} (мкМ/мин/мг)	29.6/38.9	<9.6/<9.6	<9.6/<9.6	50.3/39.6
Caco-2 проницаемость, (10^{-6} см/с)	0.5	1.6	0.4	0.3
Растворимость в воде, pH 7.4 (мкМ)	164	156	158	55
CYP P450 1A2/2C9/2C19/2D6/3A4, IC_{50} (мкМ)	24.3/15.6/9.5 /6.5/1.45	>50/36.6/40.4 /44.5/37.8	>50/>50/>50 />50/>50	0.08/24.0/13.2 /22.7/5.8
Стабильность в плазме крови человека, ($t_{1/2}$, мин)	>289.1	>289.1	>289.1	9
Связывание с белками плазмы крови, % (человек/мышь)	29.4/16.2	18.0/10.4	29.8/26.0	63.1/30.7
Противовирусная активность, SARS-CoV-2, Корейский вариант, клеточная линия Vero, EC_{50} (мкМ)	2.23	1.08	4.25	—
Цитотоксичность в клеточной линии Vero, CC_{50} (мкМ)	>50	>50	>50	—

Часть соединений-лидеров была охарактеризована в клеточном эссее на предмет наличия противовирусной активности. Клетки Vero инфицировались коронавирусом SARS-CoV-2, а

затем обрабатывались соединениями с потенциальной противовирусной активностью. Для всех протестированных соединений (**INSCoV-614-R**, **INSCoV-517D-R** и **INSCoV-614A-R**) наблюдалась заметная способность подавлять SARS-CoV-2 (см. табл. 29), сопоставимая с другими известными противовирусными агентами, такими как нирматрелвир (EC_{50} = 4.00 мкМ) и ремдесивир (EC_{50} = 7.86 мкМ), испытанных в тех же условиях. Важно отметить, что ни одно из соединений не проявляло выраженной цитотоксичности (CC_{50}) в клеточной линии Vero.

Два лучших соединения-лидера были охарактеризованы с фармакокинетической точки зрения в мышах (см. табл. 31). Пероральная биодоступность (F, %) соединения **INSCoV-517D-R** была близка к 100%, что указывало на большой потенциал этого соединения для дальнейшей оптимизации соединения-лидера. Стоит заметить, что с точки зрения структуры соединение **INSCoV-517D-R** не претерпело значительных структурных изменений на пути к клиническому кандидату **ISM-3312**.

Таблица 31. Фармакокинетический профиль избранных соединений-лидеров

Соединение	**INSCoV-614-R**	**INSCoV-517D-R**
Доза (iv/po, мг/кг)	5/20	1/20
$t_{1/2}$ [iv, ч]	0.21 ± 0.03	0.64 ± 0.15
$t_{1/2}$ [po, ч]	0.80 ± 0.09	1.07 ± 0.18
AUC_{INF} (iv, нг*ч/мл)	646 ± 120	97.8 ± 4.5
AUC_{INF} (po, нг*ч/мл)	232 ± 103	1952 ± 269
F [%]	9	99.8

Отдельно хочется подчеркнуть возможности модуля ReRSA по моделированию синтетической доступности молекулярных структур описанной кампании. Для примера рассмотрим предложенные модулем ReRSA КДИС для соединения **INCoV-601I** (см. рис. 44). Фрагменты молекулярной структуры полностью конвертированы в КДИС благодаря наличию в арсенале алгоритма трансформа Уги (см. табл. 3, R25), подразумевающего двухстадийное превращение: (1) получение изоционата из амина; (2) четырехкомпонентная реакция Уги с полученным изоционатом, амином, кетоном или альдегидом и кислотой. Биарильный фрагмент же подвергается трансформу Сузуки (табл. 3, R13.1). Таким образом, значение функции ReRSA для соединения **INCoV-601I** соответствовало 1.73, что указывало на высокую вероятность синтезируемости молекулярной структуры. Показательно сравнение КДИС, предложенных алгоритмом ReRSA, и КДИС, позже использованных в реальном синтезе соединения **INCoV-601I** (см. схему X), проведенном в CRO.

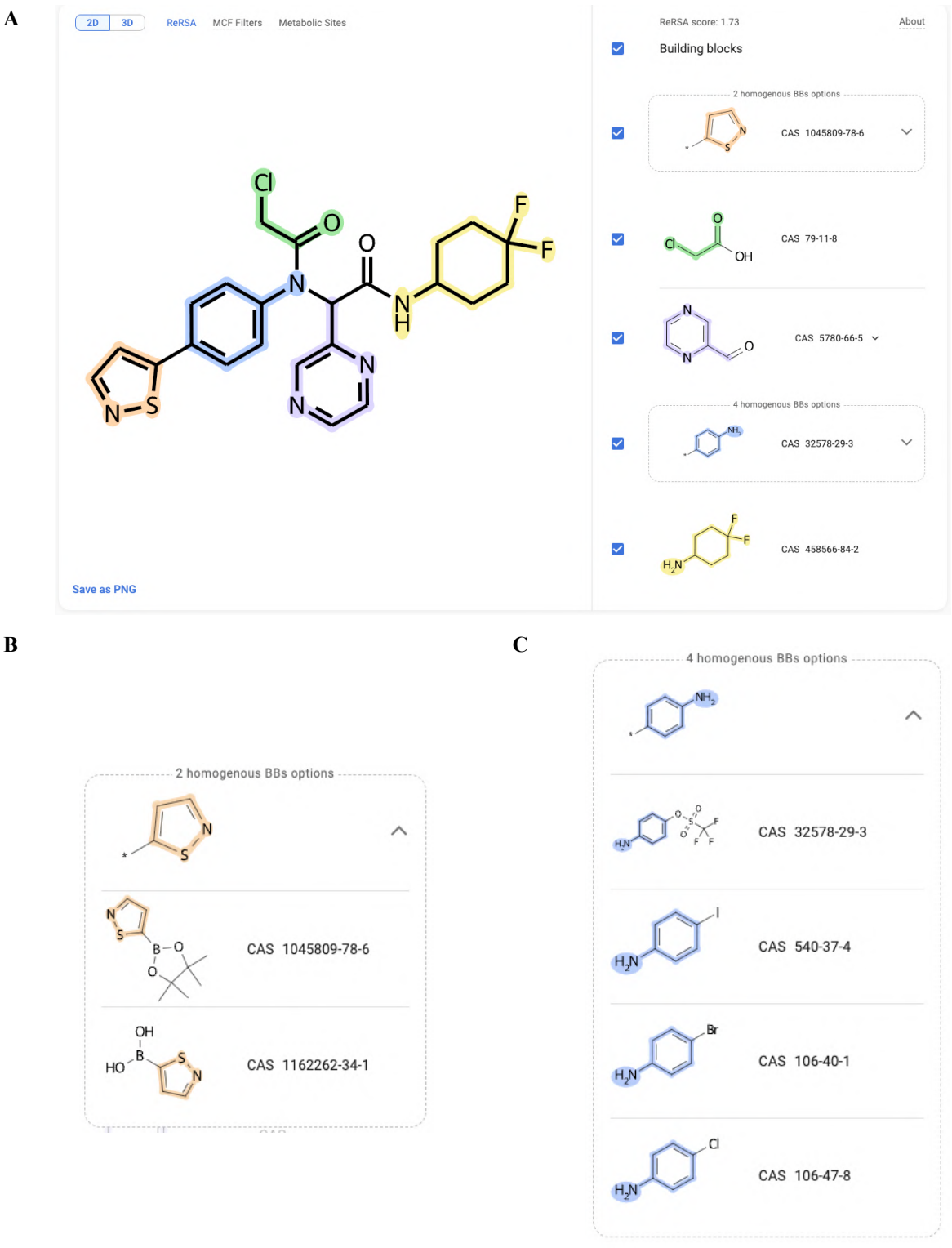


Рисунок 44. Скриншот интерфейса платформы Chemistry42. **A.** Результат работы модуля ReRSA для соединения **INCoV-601I**. **B.** Возможные исходные соединения для изотиазолового фрагмента. **C.** Возможные исходные соединения для анилинового фрагмента.

По нажатию на иконку с гомогенными строительными блоками, пользователю доступен выпадающий список с альтернативами КДИС (см. разд. 3.2.5): пинаколовый эфир изотиазол-5-бороновой кислоты **CAS 1045809-78-6** и изотиазол-5-бороновая кислота **CAS 1162262-34-1** для подсвеченного оранжевым фрагмента (рис. 44. B); пара-хлоранилин **CAS 106-47-8**, пара-броманилин **CAS 106-40-1**, пара-иоданилин **CAS 540-37-4** и пара-(трифторметилсульфонилокси)анилин **CAS 32578-29-3** — для фрагмента, подсвеченного синим цветом (рис. 44 C). Таким образом, из предложенных ReRSA КДИС можно извлечь полностью соответствующие реально использованным за исключением 1,1-дифторо-4-изоцианоциклогексана (подсвечен желтым на схеме 1), который легко может быть получен из 1,1-дифторо-4-аминоциклогексана в одну стадию, что и подразумевается трансформом Уги, закодированном в алгоритме.

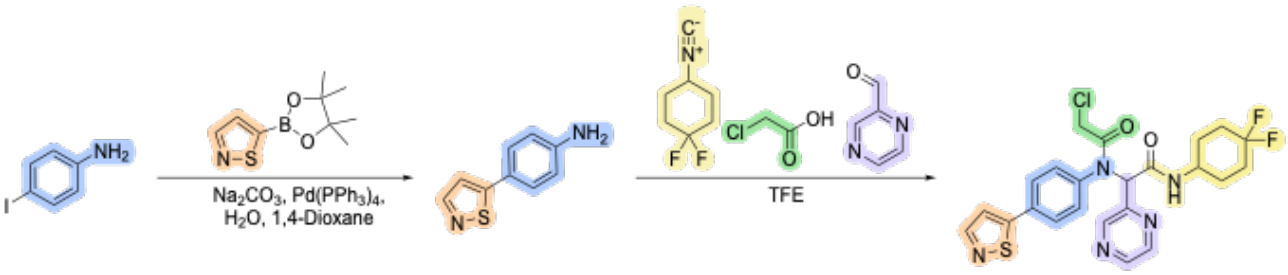


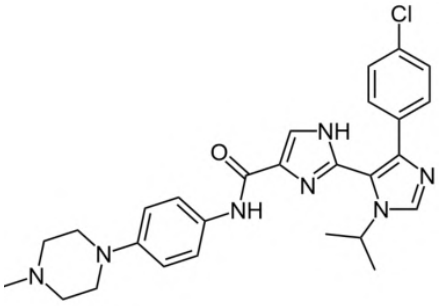
Схема 1. Схема синтеза соединения **INCoV-601I**, осуществленная в лаборатории CRO. Цветовая подсветка соответствует цветовой схеме модуля ReRSA (см. рис. 44).

Так, благодаря визуализации модулем ReRSA представляется возможным не только абстрактно оценить вероятность успешного синтеза, но и предположить схему синтезу и необходимые КДИС (больше примеров см. в разд. 3.2.3), что является крайне полезным функционалом платформы при отборе сгенерированных молекулярных структур на синтез.

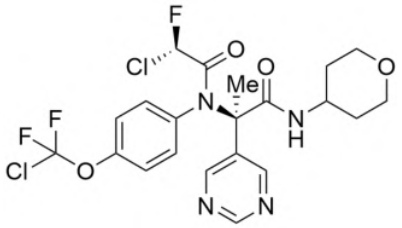
4.1.4.3 Клинические кандидаты, разработанные с применением платформы Chemistry42

Наконец, именно используя функционал сначала прототип Chemistry42, а затем и полноценную пользовательскую версию, были разработаны ведущие клинические кандидаты, находящиеся в портфеле группы компаний Insilico Medicine. В их числе ингибитор TNIK киназы **INS018_055**, успешно прошедший IIa фазу клинических испытаний в Новой Зеландии (регистрационный номер КИ: NCT05938920) в качестве терапевтического агента в лечении идиопатического фиброза легких [49], ингибитор PHD1/2 **ISM012-042**, прошедший I фазу клинических испытаний (регистрационный номер КИ: NCT06012578) [157], а также **ISM-3312**

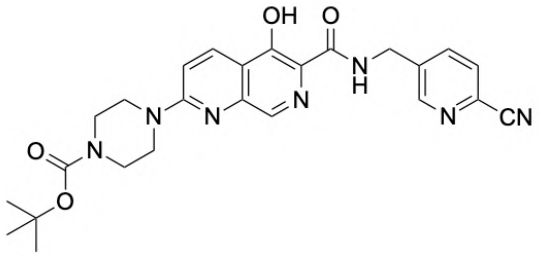
— ингибитор главной протеазы (M^{pro}) коронавируса SARS-CoV-2[12] [158] (см. рис. 45), для которого завершена I фаза клинических испытаний на здоровых добровольцах в Китае (регистрационный номер КИ: CTR20230768) в качестве однокомпонентной терапии COVID-19 [159].



INS018_055, ингибитор TNIK



ISM-3312, ингибитор Mpro SARS-CoV-2



ISM012-042, ингибитор PHD1/2

Рисунок 45. Клинические кандидаты Insilico Medicine, разработанные при помощи платформы генеративной химии Chemistry42.

[12] Кампанией по ранней разработке ингибитора M^{pro} SARS-CoV-2 руководил автор настоящей диссертации (см. разд. 3.1.4.2) — с точки зрения структуры, соединение-лидер **INSCoV-517D-R** не претерпело значительных структурных изменений на пути к клиническому кандидату **ISM-3312**.

3.2 Метод моделирования синтетической доступности ReRSA

3.2.1 Эволюция метода ReRSA

За время своего существования ММСД ReRSA претерпел значительную эволюцию (см. табл. 32), как в плане совершенствования алгоритма, так и с точки зрения пользовательского интерфейса. Первая версия ММСД ReRSA была предложена в марте 2020 года. После анализа ранее опубликованных методов моделирования синтетической доступности в качестве основы для нового метода была взята базовая гипотеза метода SA Score о влиянии встречаемости фрагментов из референсных данных на оценку синтезируемости молекул. В качестве статистического контекста использовалась информация из актуальной на март 2020 года базы данных биологически-активных малых органических молекул ChEMBL. При этом, базовым алгоритмом квази-ретросинтетической фрагментации являлся базовый метод BRICSDecompose с расширенной библиотекой робастных реакций. Например, были добавлены квази-ретросинтетические разбиения, соответствующие реакциям Сузуки и Соногаширы, синтезу сульфонамидов из аминов и сульфоацилирующих агентов и другие реакциям. Итоговое количество реакций, кодируемое в виде BRICS-подобных реакций, составило 20 для первой версии алгоритма ReRSA.

Спустя два года метод ReRSA был подвергнут глубокой переработке. В первую очередь изменения коснулись базового метода квази-ретросинтетической фрагментации, которая теперь учитывала независимость ветвей ретросинтетического древа, тогда как каждая ветвь теперь формировала самодостаточный набор синтоноподобных фрагментов — сплит. Для этого базовый метод BRICSDecompose был серьезно переработан. Помимо этого, в целях более полного охвата пространства важных для медицинской химии реакций, были добавлены те реакции, которые не нуждаются в методе BRICSDecompose, а выполняются через RunReactants (R21-R28, R38-R42 см. табл. 3). Таким образом, количество робастных реакций было увеличено почти вдвое (до 37). Принципиальным отличием от первой версии метода явилась и возможность учёта контекста коммерчески доступных стартовых соединений (КДИС). Была проработана логика конвертации синтоноподобных фрагментов в синтетические эквиваленты, которые затем индексируются по базе КДИС, что серьезным образом влияет на конечные значения оценки синтезируемости по методу ReRSA. Референсное химическое пространство было расширено почти в 3 раза, путем включения скрининг-датасетов Enamine объемом ~2M синтетически-доступных молекул. Заключительным нововведением для второй версии метода стала возможность переключения между двумя режимами (“политиками”) работы алгоритма. “Жёсткая” политика обеспечивает

фокус на синтезируемости молекулярных структур за счет критического штрафования (ReRSA = 10.0) синтоноподобных фрагментов, которые а) не индексируются в референсной базе синтоноподобных фрагментов, и б) синтетические эквивалентны которых не индексируются в базе КДИС.

Таблица 32. История развития метода ReRSA, доступного в виде модуля оценки синтезируемости в рамках платформы Chemistry42

Версия **Фактор**	**ReRSA 1.0** [89] **(2020)**	**ReRSA 2.0 (2022)**	**ReRSA 3.0 (2023)**
Референсное химическое пространство	Статистический контекст базы данных биологически активных веществ ChEMBL (~1M)	**Улучшенный статистический контекст**, дополненный базой Enamine (итого ~3M)	*Значительных изменений внесено не было*
Алгоритм квази-ретро-синтетической фрагментации	Разбиение молекулярных структур на синтоноподобные фрагменты происходит без сохранения отдельных ветвей квази-ретро-синтетического древа	Разбиение молекулярных структур на фрагменты происходит с учетом отдельных ветвей квази-ретро-синтетического древа образующих **сплиты**	
Контекст исходных соединений	Контекст не учитывается	Контекст датасета КДИС (~200K) значительно влияет на оценку синтезируемости молекулярных структур	Все КДИС были аннотированы идентификационным и номерами **CAS ID**. Стала доступна **визуализация** результатов работы модуля.

Продолжение на следующей странице

Продолжение таблицы 32

Версия **Фактор**	**ReRSA 1.0** [89] **(2020)**	**ReRSA 2.0 (2022)**	**ReRSA 3.0 (2023)**
Контекст бизнес-логики дизайна потенциальных лекарственных веществ	Неиндексируемые в референсном датасете синтоноподобные фрагменты не получают критический штраф.	Доступны две политики оценки синтезируемости исходя из бизнес-логики: “мягкая” с фокусом на новизну молекулярных структур и “жёсткая” политика с фокусом на синтетическую доступность, обеспечивающая критический штраф для неиндексируемых фрагментов.	Доступна дополнительная быстрая фильтрация синтетически нерелевантных **5-членных ароматических гетероциклов** в рамках “жесткой” политики.
Робастные реакции	Используется 20 реакций	Используется 37 реакций	Используется **52** реакции, включая реакции **макроциклизации** и **гетероциклизации**.

“Мягкая” политика, которая использовалась по умолчанию в версии ReRSA 1.0, теперь стала опциональной и используется в целях выхода за пределы референсируемого химического пространства за счет некритичных штрафов в итоговое значение оценки синтезируемости.

Главной целью развития третьей (на данный момент последней) версии ММСД ReRSA стало предоставление возможности пользователю платформы генеративной химии интерпретировать результаты работы алгоритма оценки синтезируемости путем визуального анализа предоставляемого набора найденных алгоритмом КДИС. Это потребовало унификации идентификационных номеров КДИС, путем перехода на общепринятую систему индексации химических веществ CAS. Так же это потребовало проработки функций отрисовки молекулярных структур и подсветки релевантных КДИС частей этих молекулярных структур. Особую важность в последнем представляло создание чёткой системы ссылок между синтетическими эквивалентами и соединениями из базы КДИС (см. разд. 2.2.7). Пользовательский интерфейс, разработанный для визуализации КДИС позволяет копировать CAS ID и SMILES-строки предлагаемых алгоритмом исходных соединений (см. рис. 46).

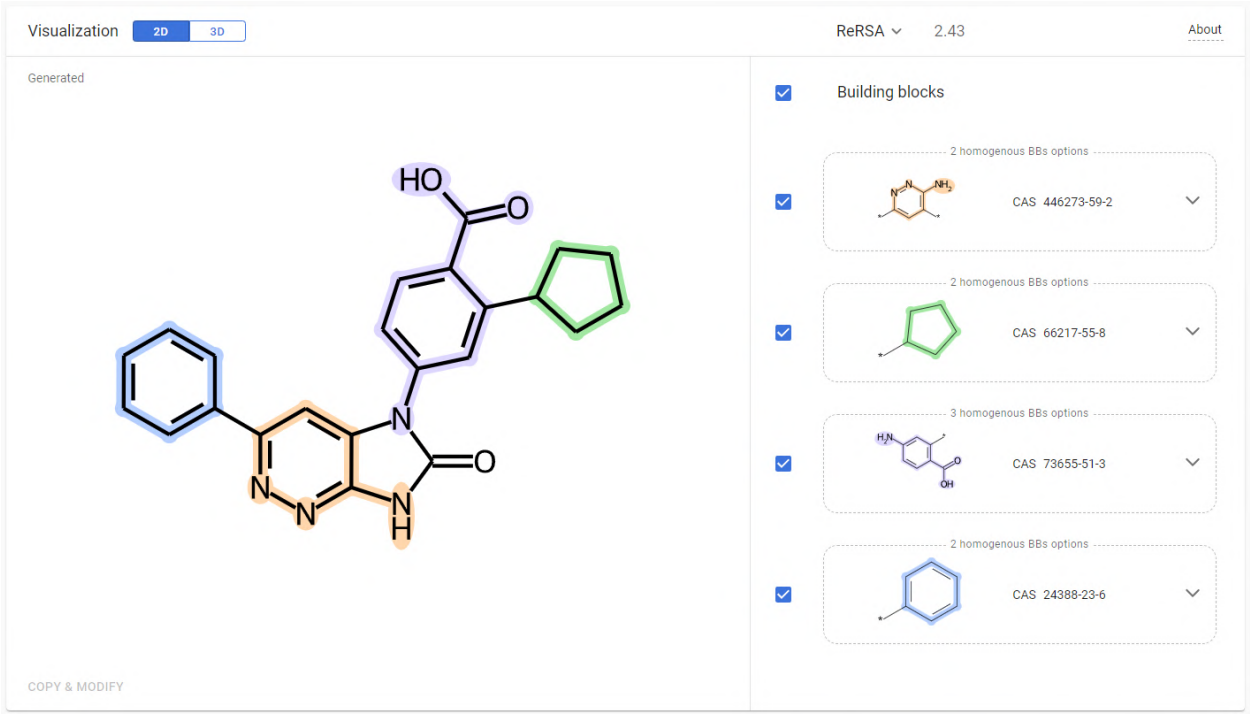


Рисунок 46. Пользовательский интерфейс визуализатора ReRSA, представленный на платформе Chemistry42.

При этом предоставляется возможность увидеть не одно КДИС, для каждой из частей молекулярной структуры, сконвертированной соответствующим образом, а все найденные опции (*homologous BBs options*, гомологичные варианты исходных соединений) для выбора пользователем наиболее релевантных его задачам КДИС. Так, например, для фрагмента с оранжевой подсветкой в качестве исходных соединений предлагаются на выбор коммерчески доступные 3-амино-4-бром-6-хлорпиридазин (**CAS 446273-59-2**) и 3-амино-4,6-дихлорпиридазин (**CAS 1161847-28-4**) (см. рис. 47). А для фрагмента молекулярной структуры с сиреневой подсветкой в качестве исходных соединений предлагаются на выбор коммерчески доступные циклопентил-борпинаколат **CAS 66217-55-8**) и циклопентилбороновая кислота (**CAS 63076-51-7**). В зависимости от персональных предпочтений, опыта работы и реконструируемой схемы синтеза, в том числе, по соображениям региоселективноти, пользователь может выбрать тот или иной вариант КДИС из перечня предлагаемых.

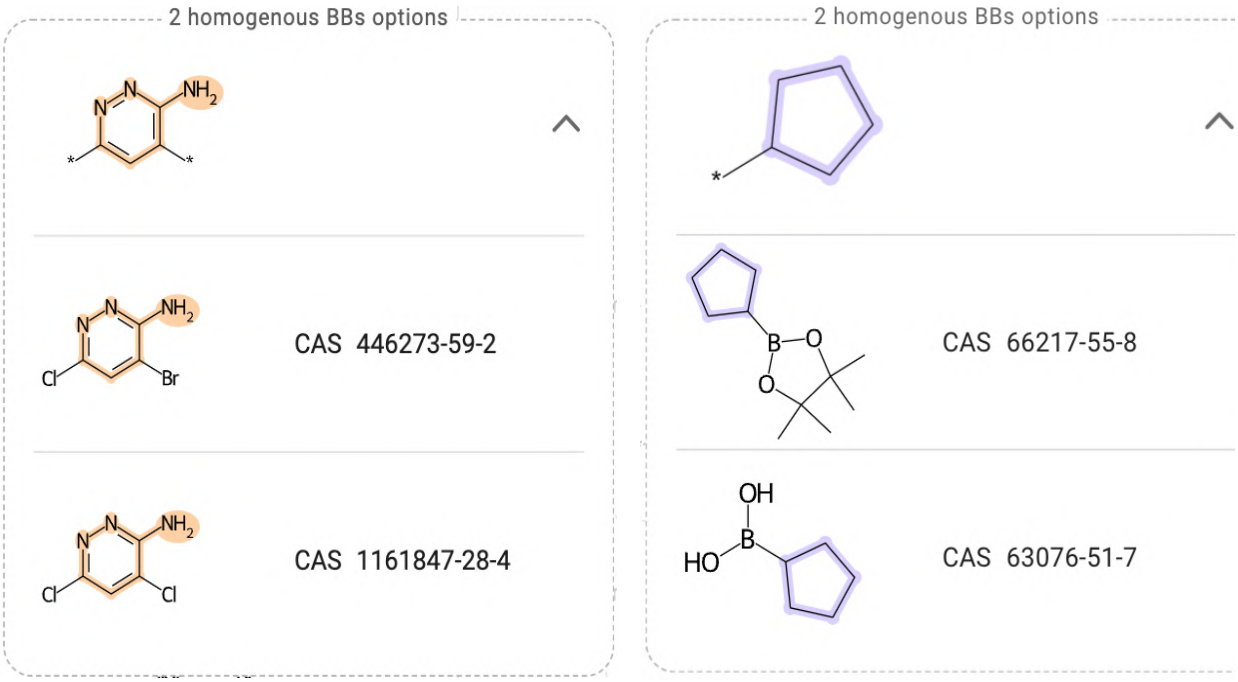


Рисунок 47. Гомологичные варианты КДИС, предлагаемые алгоритмом ReRSA на платформе Chemistry42.

Параллельно задачам, связанным с визуализацией, требовалось улучшить часть алгоритма, ответственного за фильтрацию синтетически нерелевантных структур. В ходе работы с предыдущими версиями алгоритма и платформы генеративной химии от пользователей платформы регулярно поступала обратная связь, которая сводилась к скептическому взгляду относительно предлагаемых алгоритмом ReRSA оценкам синтезируемости молекулярных структур, содержащих пятичленные ароматические гетероциклы, которые требовалось дополнительно проверять в референсных базах на предмет их прецедентности. В связи с этим было выдвинуто предложение расширить базовый алгоритм ReRSA за счёт дополнительной фильтрации подструктур, соответствующих синтетически нерелевантным 5-членным ароматическим гетероциклам (см. разд. 2.2.9, 3.2.2.2 и 3.2.3.7) в рамках “жесткой” политики алгоритма. По материалам разработки этого дополнительного фильтрационного алгоритма была подготовлена и успешно защищена в стенах Химического факультета МГУ имени М. В. Ломоносова дипломная работа, за авторством Бондарева Никиты и научным руководством автора настоящей диссертации[13]. Также результаты разработки были суммированы в научной публикации [160]. Стоит отметить, что помимо

[13] Бондарев Н., ВКР “Прогнозирование синтетической доступности потенциальных лекарственных веществ, содержащих гетероциклические фрагменты”. Науч. рук.: доц., к.х.н. Радченко Е. В., асп. Загрибельный Б. А. Химический факультет МГУ, Москва, 2023.

фильтрационной составляющей алгоритма описанное исследование инициировало внедрение дополнительного блока реакций, приводящих к синтезу наиболее распространенных 5-членных ароматических гетероциклов (R29-R37, см. табл. 3). Заключительным нововведением в третьей версии модуля ReRSA стала синхронизация алгоритма с генерацией макроциклических структур. И хотя на данный момент на платформе в рамках рабочих сценариев и модельных экспериментов недоступна макроциклизация в виде удобного и универсального процесса, тем не менее модуль ReRSA может адекватно оценивать макроциклические структуры на предмет их синтезируемости за счет внедрения в общий пул реакций, соответствующих наиболее распространенным методам макроциклизации (R43-R52, см. табл. 3).

3.2.2 Валидация метода ReRSA на зарегистрированных лекарственных веществах и синтезированном химическом пространстве

3.2.2.1 Валидация квази-ретросинтетической компоненты модуля ReRSA на зарегистрированных лекарственных веществах и клинических кандидатах

Исходя из обсуждаемой ранее (см. разд. 1.4.3) гипотезы о том, что ретросинтетический анализ является лучшим методом моделирования синтетической доступности, представляется, что надежным методом ретроспективной *in silico* валидации метода оценки синтезируемости молекулы является сравнение (в той мере, в которой это возможно) результатов работы алгоритма оценки синтезируемости с реальной синтетической схемой ранее полученных соединений. В качестве подобных соединений можно использовать зарегистрированные лекарственные вещества, поскольку именно они представляют интерес с точки зрения фокуса работы платформы генеративной химии Chemistry42, и схемы их синтеза можно легко найти в профильной литературе.

В целях независимого исследования 4 экспертам в области медицинской и органической химии, работающих в группе компаний Insilico Medicine, был предоставлен список из 400 зарегистрированных за последние 15 лет лекарственных веществ и клинических кандидатов, находящихся сейчас на последних фазах клинических исследований. Список был разделен на 4 равные части и экспертам было предложено выбрать 20 молекул из 100 для ретроспективной валидации метода ReRSA. Итоговый список из 80 молекул представлен в таблице 33.

Основным критерием, который влиял на оценку для каждого из лекарственных веществ, была способность конвертировать молекулярную структуру вещества в КДИС.

Успешным выполнением процедуры считалась полная конвертация в КДИС. Помимо этого, внимание уделялось и тому, насколько предложенные методом ReRSA КДИС соответствуют тем исходным соединениям, которые были использованы для получения конечных веществ в рамках опубликованных ранее синтетических схем.

По итогам *in silico* эксперимента полная конверсия молекулярных структур в КДИС наблюдалась для 45 соединений из 80. При этом для большей части соединений, которые не были полностью сконвертированы, степень конверсии оставалась высокой, то есть не были сконвертированы в КДИС единичные фрагменты, чаще всего циклической природы. В целях иллюстрации результатов валидационного эксперимента конвертация молекулярных структур соединений с указанием идентификационных номеров CAS КДИС из таблицы сопоставлялась со структурами исходных соединений из опубликованных схем синтеза с консистентной по цвету подсветкой подструктур эквиретросинтетического происхождения.

Таблица 33. Зарегистрированные лекарственные препараты, отобранные экспертами для валидации алгоритма ReRSA (в столбцах “ПК” отмечена полная конверсия фрагментов в КДИС, зеленая заливка ячеек соответствует полной конверсии указанной структуры)

№	Название	ПК	ReRSA	№	Название	ПК	ReRSA
1	(S)-циталопрам ((S)-citalopram)	да	2.85	41	омариглиптин (omarigliptin)	нет	7.16
2	апалутамид (apalutamide)	нет	4.28	42	пазопаниб (pazopanib)	да	2.61
3	апатиниб (apatinib)	да	2.42	43	пакритиниб (pacritinib)	да	3.51
4	апиксабан (apixaban)	нет	4.33	44	пемигатиниб (pemigatinib)	нет	4.49
5	асунапревир (asunaprevir)	да	2.92	45	понатиниб (ponatinib)	да	2.66
6	аталурен (ataluren)	да	1.80	46	радотиниб (radotinib)	да	2.50
7	балоксавир (baloxavir)	нет	6.05	47	релуголикс (relugolix)	нет	4.45
8	барицитиниб (baricitinib)	да	2.08	48	рибоциклиб (ribociclib)	да	2.41
9	беклабувир (beclabuvir)	нет	4.92	49	ривароксабан (rivaroxaban)	да	2.25
10	беротралстат (berotralstat)	нет	5.02	50	римегепант (rimegepant)	нет	4.91
11	биктегравир (bictegravir)	нет	6.74	51	ритлецитиниб (ritlecitinib)	да	2.32
12	бригатиниб (brigatinib)	да	2.48	52	рофлумиласт (roflumilast)	да	2.20
13	ванипревир (vaniprevir)	да	4.65	53	руксолитиниб (ruxolitinib)	да	2.09
14	вемурафениб (vemurafenib)	нет	3.75	54	сакубитрил (sacubitril)	да	2.17
15	гемиглиптин (gemigliptin)	нет	4.84	55	сафинамид (safinamide)	да	1.74
16	гепотидацин (gepotidacin)	нет	5.48	56	селинексор (selinexor)	нет	3.88

Продолжение на следующей странице

Продолжение таблицы 33

№	Название	ПК	ReRSA	№	Название	ПК	ReRSA
17	гилтеритиниб (gilteritinib)	нет	3.86	57	селперкатиниб (selpercatinib)	нет	4.50
18	гласдегиб (glasdegib)	да	2.49	58	сорафениб (sorafenib)	да	2.21
19	данопревир (danoprevir)	нет	6.48	59	суворексант (suvorexant)	да	2.69
20	даридорексан (daridorexant)	да	2.69	60	таземетостат (tazemetostat)	да	2.57
21	деукравацитиниб (deucravacitinib)	да	2.38	61	тезакафтор (tezacaftor)	нет	5.39
22	ивосидениб (ivosidenib)	да	2.04	62	телапревир (telaprevir)	да	2.82
23	икотиниб (icotinib)	да	3.65	63	телотристат этил (telotristat ethyl)	да	2.57
24	карфилзомиб (carfilzomib)	да	2.64	64	тофацитиниб (tofacitinib)	да	2.20
25	кобицистат (cobicistat)	да	2.79	65	трофинетид (trofinetide)	да	1.97
26	копанлисиб (copanlisib)	нет	4.56	66	уброгепант (ubrogepant)	нет	6.60
27	кризотиниб (crizotinib)	да	2.65	67	уденафил (udenafil)	да	2.83
28	ларотректиниб (larotrectinib)	да	2.62	68	фенебрутиниб (fenebrutinib)	нет	5.14
29	лемборексан (lemborexant)	нет	5.09	69	флуматиниб (flumatinib)	да	2.62
30	лениолисиб (leniolisib)	да	2.51	70	формотерол (formoterol)	да	2.03
31	лифитеграст (lifitegrast)	нет	4.87	71	эвоглиптин (evogliptin)	нет	4.05
32	лорлатиниб (lorlatinib)	нет	6.76	72	элбасвир (elbasvir)	нет	5.35
33	лотиланер (lotilaner)	нет	5.94	73	элобиксибат (elobixibat)	нет	4.77
34	маравирок (maraviroc)	да	2.93	74	элуксадолин (eluxadoline)	нет	5.57
35	нирогакестат (nirogacestat)	нет	5.91	75	энародустат (enarodustat)	нет	4.61
36	обицетрапиб (obicetrapib)	нет	4.90	76	энзалутамид (enzalutamide)	да	2.20
37	озанимод (ozanimod)	да	2.37	77	энкорафениб (encorafenib)	нет	3.12
38	озеноксацин (ozenoxacin)	нет	4.79	78	энсартиниб (ensartinib)	да	2.90
39	олапариб (olaparib)	да	2.50	79	эфавиренз (efavirenz)	да	2.46
40	олицеридин (oliceridine)	нет	5.02	80	эфиноконазол (eficonazole)	да	1.99

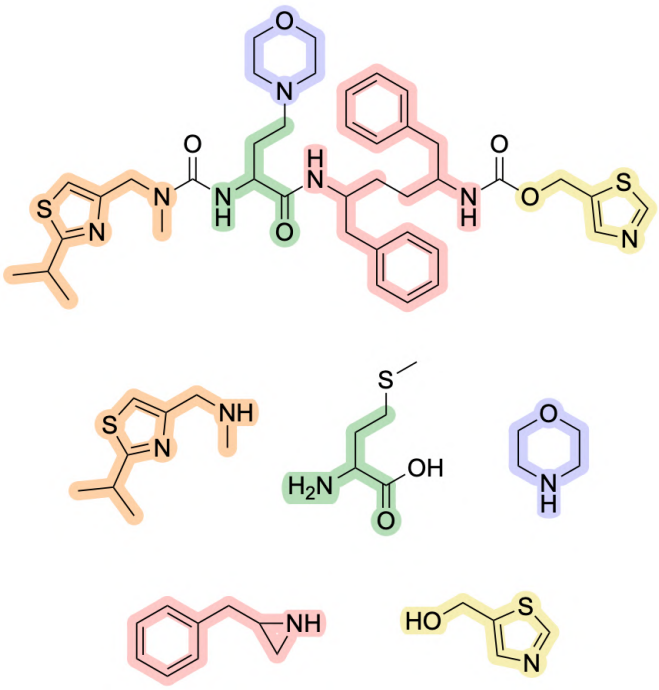
Молекулярная структура кобицистата (табл. 33, №25), ингибитора цитохрома CYP3A4, применяемого в качестве компонента ВААРТ (высокоактивной антиретровирусной терапии) ВИЧ-инфекции, была успешно фрагментирована при помощи квази-ретросинтетического алгоритма ReRSA в коммерчески доступные исходные соединения (КДИС) (см. рис. 48). Если сравнить полученные КДИС с исходными соединениями, которые предложены в оригинальной схеме синтеза кобицистата [161], то можно обнаружить несколько идентичных стартовых соединений (морфолин **CAS 110-91-8** и 5-гидроксиметилтиазол **CAS 38585-74-9**), или очень похожих на них (амин **CAS 903131-67-9** и производный от аспарагиновой кислоты альдегид **CAS 498-20-4**). Ретросинтетическое отличие между двумя результатами состоит в наборе реакций, который используется в ходе ретросинтеза и в критерии остановке

ретросинтетических разбиений. Так, например, гомологичное амину **CAS 903131-67-9** исходное соединений из оригинальной схемы синтеза не может быть получено по методу ReRSA, поскольку метод ReRSA выполняет разбиение всех связей, соответствующих реакциям из таблицы 3, в том числе и связь, соответствующую N-алкилированию (R3), в результате чего метод ReRSA предлагает в качестве КДИС йодистый метил **CAS 74-88-4**. Использование методом ReRSA альдегида **CA 498-20-4** для связывания с морфолином **CAS 110-91-8,** обуславливается наличием трансформа восстановительного аминирования (R15), в то время как в оригинальном синтезе используется нестандартное нуклеофильное замещение вторичной аминогруппой морфолина метилсульфидной уходящей группы метионина. Различие в исходных соединениях наблюдается так же и для 1,6-дифенилгексан-2,5-диаминового фрагмента (светло-красная подсветка на рис. 48). В то время как алгоритм ReRSA обнаружил непосредственно сам 1,6-дифенилгексан-2,5-диамин в базе КДИС (**CAS 144186-34-5**), оригинальная схема синтеза сводится к более глубокой конверсии этой части молекулярной структуры в 2-бензилазиридин. При этом мы не считаем данное различие принципиальным и влияющим на оценку синтезируемости молекулярной структуры, пока не заданы специфические критерии и условия (см. 1.4.2), согласно которым один набор КДИС будет более предпочтителен, чем другой.

В качестве другого примера молекулярная структура ванипревира (табл. 33, №13), макроциклического ингибитора NS3/4A протеазы вируса гепатита человека, была успешно фрагментирована при помощи квази-ретросинтетического алгоритма ReRSA в коммерчески доступные исходные соединения (КДИС) (см. рис. 49). Обратим внимание на то, что до версии ReRSA 3.0 не было возможности подвергать квази-ретросинтетическому анализу макроциклические структуры. Аналогично с кобицистатом некоторые части молекулярной структуры ванипревира были сконвертированы алгоритмом ReRSA в КДИС, способом аналогичным той логике, что применяется в оригинальной схеме синтеза [162]. Так, например набор КДИС полностью совпадает с оригинальными исходными соединениями в части циклопропилсульфонамида **CAS 154350-29-5** (желтая подсветка) и трет-бутиламиноуксусной кислоты **CAS 471-50-1** (розовая подсветка), и с точностью до защитной группы в отношении 3-гидроксипролина **CAS 51-35-4** (зеленая подсветка). Очевидно, близко к оригинальной схеме синтеза бром-производное изоиндолина **CAS 127168-81-4** (оранжевая подсветка), которое может быть получено в 1–2 стадии из 1-бромо-2,3-бис(бромометил)бензола, обозначенного в способе получения ванипревира. Отметим, что относительно высокое значение ReRSA для ванипревира (**5.13**), несмотря на высокую конверсию в КДИС, обусловлено штрафующей поправкой на наличие в молекулярной структуре макроцикла (см. 2.2.10). В то же время обе

схемы синтеза (одна реальная, вторая гипотетическая) сходятся одинаково к цикл-замыкающему метатезису, в качестве метода макроциклизации. Что касается циклопропанового фрагмента, то поскольку в методе ReRSA не используются реакции синтеза трёхчленного кольца, квази-ретросинтетический анализ остановился на этапе конвертации в 1-амино,2-этилциклопропановую кислоту (**CAS 87480-58-8,** светло-красная подсветка), в то время как оригинальная схема предлагает более глубокую конверсию в глицин и 1,4-дибромэт-2-ен (сиреневая подсветка).

А



Б

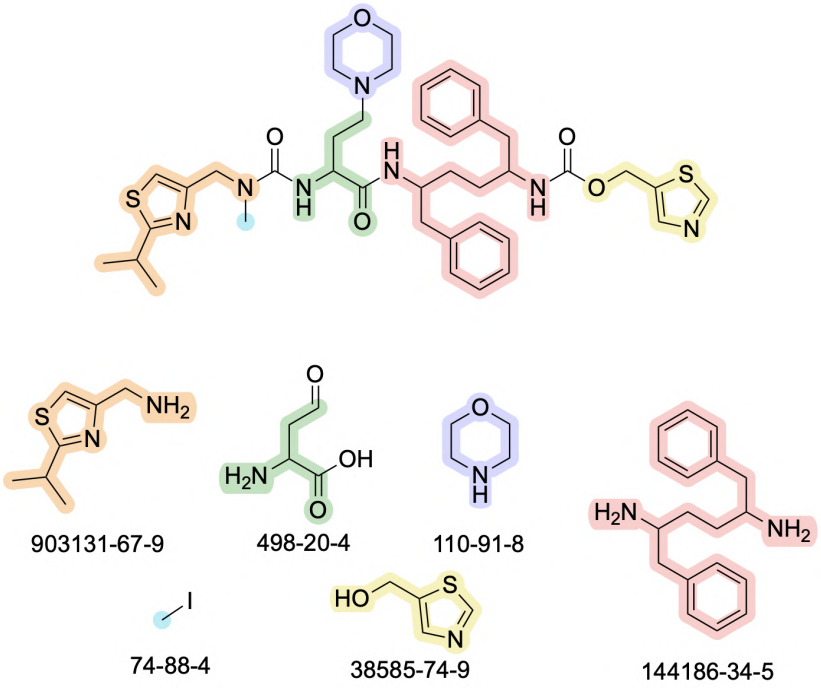
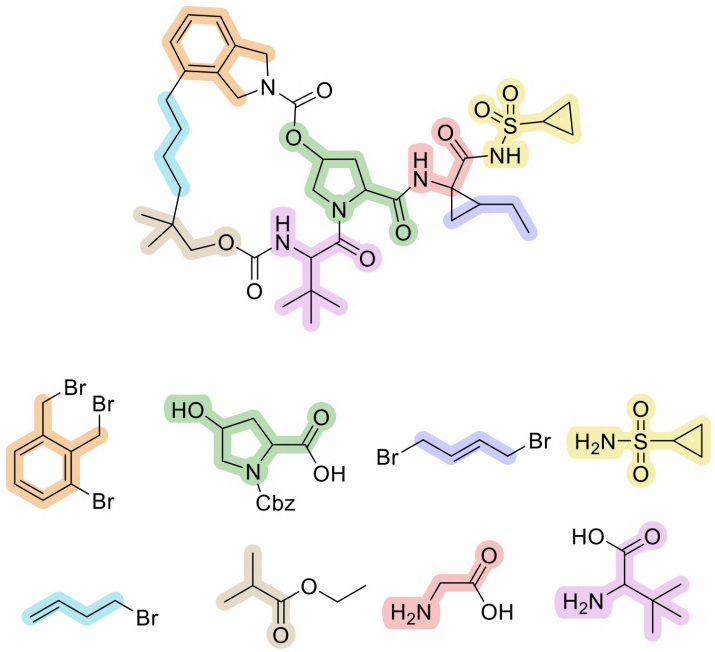


Рисунок 48. А. Исходные соединения, предлагаемые в оригинальной схеме синтеза кобицистата [161]. **Б.** КДИС, предлагаемые ММСД ReRSA в ходе квазиретросинтетического анализа кобицистата

А



Б

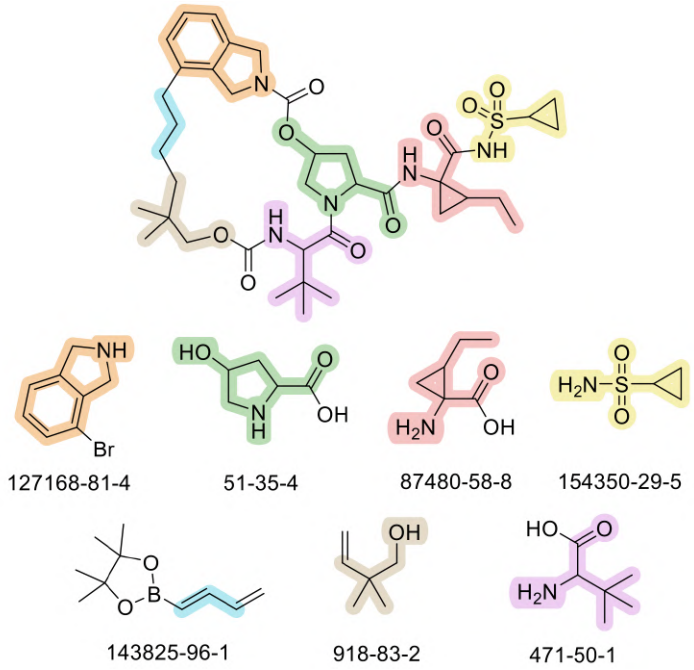


Рисунок 49. А. Исходные соединения, предлагаемые в оригинальной схеме синтеза ванипревира [162]. **Б.** КДИС, предлагаемые ММСД ReRSA в ходе квазиретросинтетического анализа ванипревира.

Картина квази-ретросинтетической фрагментации, которая практически полностью совпадает с исходными соединениями из оригинальной синтетической схемы [163], наблюдалась в случае энзалутамида, антагониста андрогенового рецептора, применяемого в терапии кастрат-резистентного рака предстательной железы. Минимальная разница между КДИС, предлагаемыми алгоритмом ReRSA и предложенными в статье исходными соединениями заключалась в том, что алгоритм провёл ретро-реакцию алкилирования первичных амидов с выходом на йодистый метил (**CAS 74-88-4**) и 4-бром,2-фторбензамид (**CAS 292621-45-5**), в то время как оригинальный синтез выходит на соответствующее производное бензойной кислоты и метиламин (см. рис 50).

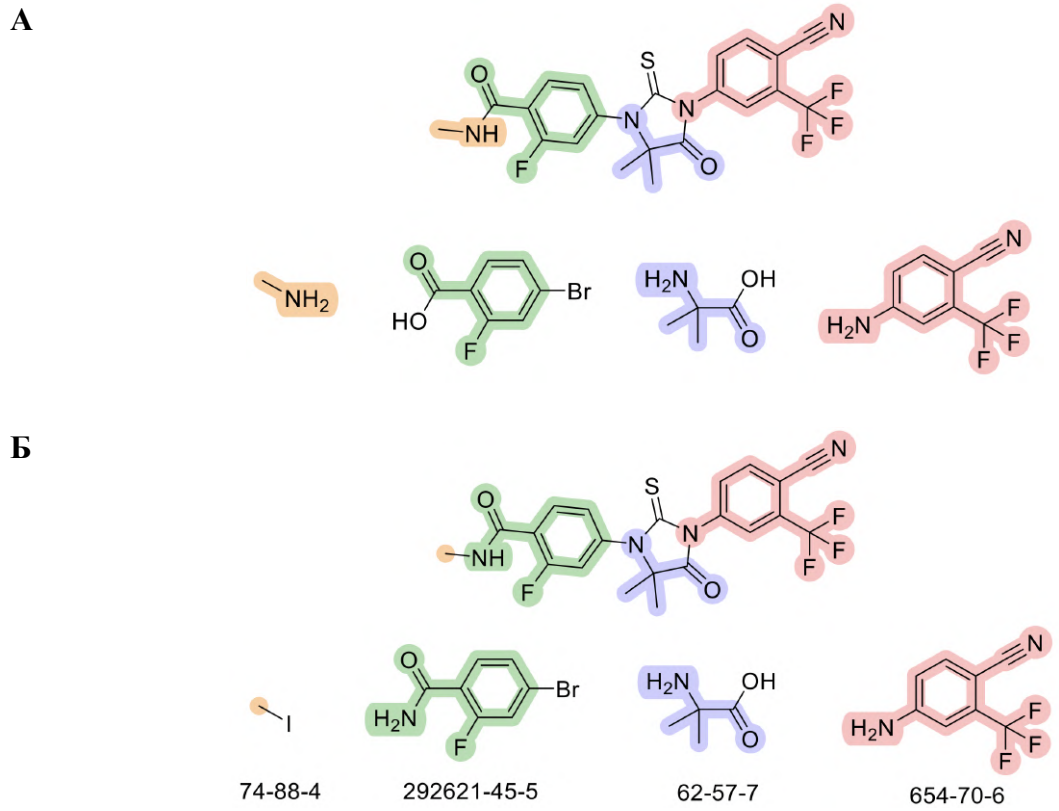


Рисунок 50. А. Исходные соединения, предлагаемые в оригинальной схеме синтеза энзалутамида [163]. **Б.** КДИС, предлагаемые ММСД ReRSA в ходе квазиретросинтетического анализа энзалутамида.

Основные проблемы алгоритма ReRSA наблюдаются в том случае, когда целевая молекулярная структура содержит подструктуры, соответствующие гетероциклам, синтезы которых не покрываются блоком реакций R29-R37 (табл. 3). Так, например, структура лотиланера содержит 4,5-дигидризооксазольный фрагмент, который синтезируется через [3+2]-диполярное циклоприсоединение, который лишь частично обрабатывается алгоритмом

ReRSA в “мягкой” политике [164]. Поскольку этот тип гетероциклов не покрывается методом ReRSA, то эта часть молекулярной структуры соединения остается не фрагментированной и соответственно не конвертируемой в синтетические эквиваленты, что в конечном итоге отражается на итоговом значении оценки синтезируемости ReRSA = 5.94 (см. рис. 51).

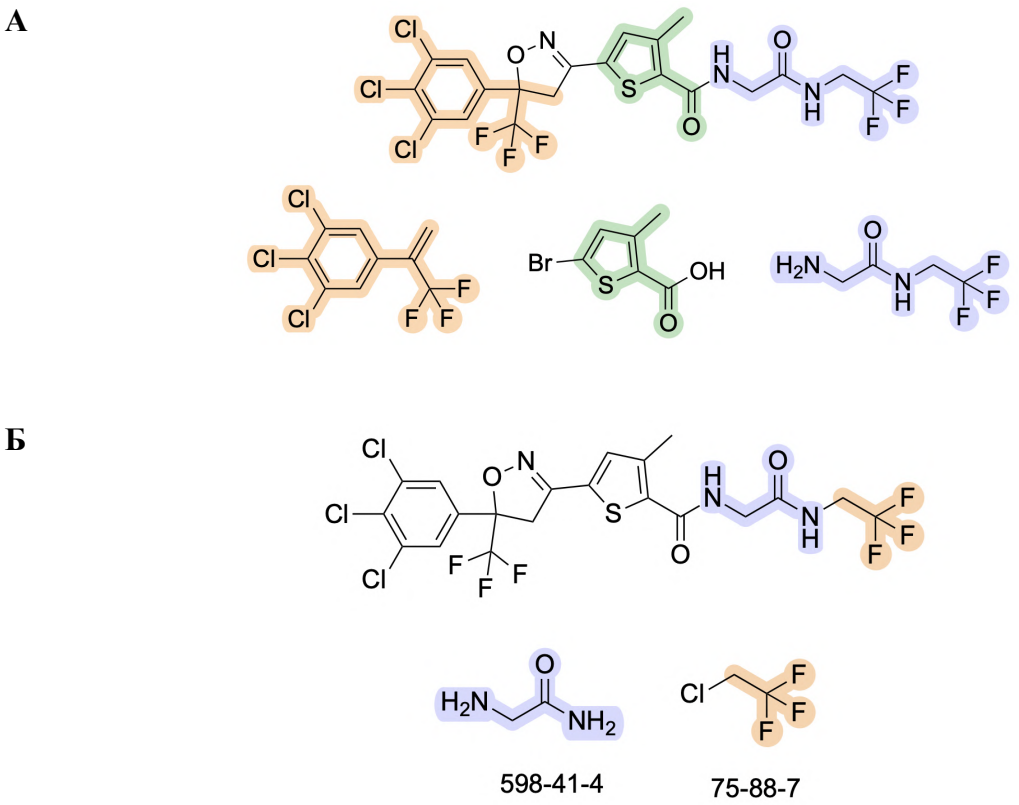


Рисунок 51. А. Исходные соединения, предлагаемые в оригинальной схеме синтеза лотиланера [164]. **Б.** КДИС, предлагаемые ММСД ReRSA в ходе квазиретросинтетического анализа лотиланера.

Справедливо отметить, что данный эксперимент является довольно жестким по отношению к любому ММСД, поскольку многие зарегистрированные лекарственные вещества синтезируются согласно длинным синтетическим схемам. Например, в синтезе лемборексанта, антагониста двух орексиновых рецепторов, применяемого в лечении бессонницы, речь идет фактически о 10 последовательных реакциях, а в синтетической схеме для уброгепанта, первого препарата в классе пероральных антагонистов рецептора пептида, связанного с геном кальцитонина (CGRP, *calcitonin gene-related peptide*), можно проследить даже 25 стадий, организованных в виде конвергентного синтеза [165]. В то же время, основной фокус платформы генеративной химии Chemistry42 лежит в поле генерации молекулярных структур, подобных соединениям-хитам (*hit-like generation*), и их первичной оптимизации, в то время как зарегистрированные лекарственные вещества представляют собой продукты

трудоемкой структурной оптимизации на этапе улучшения свойств соединения-лидера или на поздних этапах оптимизации соединений-хитов. Именно на этих более поздних этапах разработки потенциальных лекарственных веществ производятся изменения структуры с целью улучшения фармакокинетических свойств хемотипа, которые требуют удлинения и усложнения синтетических схем. По-видимому, наличие датасета, который бы содержал информацию о структурах первичных соединений-хитов, ставших началом в разработке зарегистрированных лекарственных веществ, а также структуры отдельных соединений в прогрессе кампании разработки на ранних этапах (соединения-хиты второй очереди, оптимизированные соединения-хиты), было бы крайне полезным и более релевантным для валидации ММСД, чем ретроспективный анализ синтетических схем лекарственных веществ. Тем не менее, описанный эксперимент, является в своём роде, предельным испытанием для любого ММСД, и результат в 56% (45 из 80) полностью конвертированных в КДИС молекулярных структур представляется удовлетворительным.

3.2.2.2 Валидация модуля ReRSA на предмет фильтрации синтетически нерелевантных 5-членных ароматических гетероциклов среди референсного химического пространства

Помимо этого, на базах данных синтезированных и коммерчески доступных соединений был валидирован модуль ReRSA, ответственный за фильтрацию молекулярных структур, содержащих синтетически нерелевантные подструктуры, описанный в разделе 2.2.9 (см. табл. 34).

Таблица 34. Анализ наборов коммерчески доступных молекул

Название набора	**Число молекул**	**Число отфильтрованных молекул**
Малые лекарственные молекулы, одобренные FDA в период с 2000 по 2022 г.	469	1 (0.21%)
Малые лекарственные молекулы, проходящие фазу клинических испытаний	6 981	29 (0.42%)
Каталог коммерчески доступных биоактивных молекул WuXi [166]	117 228	106 (0.09%)
Каталог коммерчески доступных биоактивных молекул Angene [167]	92 816	312 (0.34%)
Каталог коммерчески доступных биоактивных молекул ChemDiv [112]	1 541 619	17 364 (1.13%)

Единственной отсеянной молекулой среди малых лекарственных молекул, одобренных FDA, является молекула аватромбопага (рис. 52). Присутствующий в структуре аватромбопага 4-хлор-2-R-тиофен (R — ароматический заместитель) встретился в базе ChEMBL 29 лишь 8 раз.В остальных наборах синтезированных соединений также была отфильтрована часть молекул, содержащих редкие или не обнаруженные в референсных наборах подструктуры.

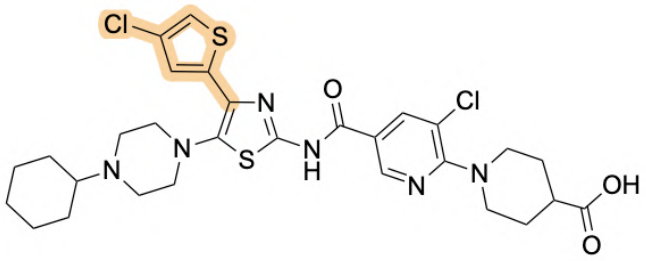


Рисунок 52. Структура аватромбопага. Подсвечена подструктура, обнаруженная в базе данных ChEMBL 29 8 раз.

Проведенные эксперименты по анализу наборов синтетически доступных, т. е. уже синтезированных, молекул являются показательными, поскольку в очередной раз подчеркивают основную проблему науки о данных — отсутствие готового всеобъемлющего обучающего набора [97]. В целом, эксперименты на выборках реальных молекул продемонстрировали, что модуль фильтрует незначительную долю молекул с действительно редко встречающимися фрагментами, что формально является ложноотрицательным срабатыванием, но допустимо с точки зрения ранее описанной бизнес-логики рационального дизайна лекарств с применением генеративных моделей.

Принимая это во внимание, разработанный алгоритм был построен именно так, чтобы обеспечивать легкую модификацию. Разработан модуль для дообучения модели, которому на вход требуется подать список синтезированных ранее молекул в виде SMILES-строк, а алгоритм в автоматическом режиме пересоберет библиотеку штрафующих SMARTS-подструктур с учетом всех оптимизаций (кластеризации однородных штрафующих фрагментов и иерархической кластеризации). Также можно изменить критерий (порог) классификации фрагментов как штрафующих, например, все фрагменты, встретившиеся в обучающей выборке хотя бы один раз, считать не фильтрующими, а подструктуры, отсутствующие в обучающей выборке, определить как синтетически нерелевантные и, соответственно, запрещенные. Алгоритм позволяет изменять указанный параметр и в автоматическом режиме пересобирать библиотеку подструктур. Предполагается дообучать

модель с некоторой периодичностью (например, раз в полгода) с использованием обновлений публично доступных химических баз, а также проприетарных данных.

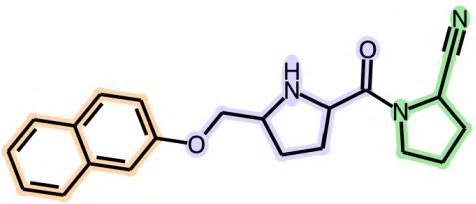
3.2.3 Валидация метода ReRSA на структурах, полученных методами генеративной химии

В качестве валидации метода ReRSA на структурах, полученных методами генеративной химии, мы анализировали избранные молекулярные структуры с наивысшими значениями интегральной функции награды и с наилучшими оценками в других ключевых 3D модулях (PLI Score, фармакофорный модуль, Shape модуль) на предмет того, насколько информация, полученная о КДИС, соответствует возможности восстановить схему синтеза для избранного соединения. Для этого валидационного эксперимента были взяты результаты ранее описанных модельных экспериментов (см. табл. 13).

3.2.3.1 Валидация модуля ReRSA на основе модельного эксперимента № 1

Результат применения метода моделирования ReRSA к избранной молекулярной структуре **INS-009923** из модельного эксперимента № 1 (см. разд. 3.1.3.1) представляет собой набор из 3 КДИС (см. рис. 53).

INS-009923, ReRSA = 2.46



CAS 180-13-2 CAS 5626-49-3 CAS 1336214-19-7

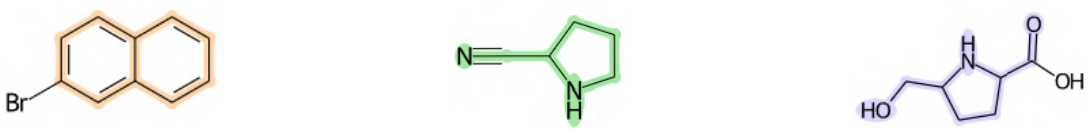


Рисунок 53. Набор КДИС, предложенный алгоритмом ReRSA для соединения **INS-009923.**

Синтетический анализ соединения **INS-009923** (см. схему 2), исходя из КДИС, предложенных методом ReRSA, может выглядеть следующим образом. Аминокислота **CAS 1336214-19-7** прежде чем вступить в реакцию с амином **CAS 5626-49-3** должна быть защищена, например при помощи Boc-защитной группы, с целью предотвращения побочной реакции ацилирования вторичной аминогруппы. Затем можно проводить реакцию амидного

синтеза в присутствии активирующего агента карбоксидиимидазола (CDI). Полученный аддукт затем можно ввести в реакцию каталитического О-арилирования с арилбромидом **CAS 180-13-2**. Образующееся вещество может быть затем обработано трифторуксусной кислотой с целью удаления Boc-защитной группы, что в итоге даст искомое соединение **INS-009923**.

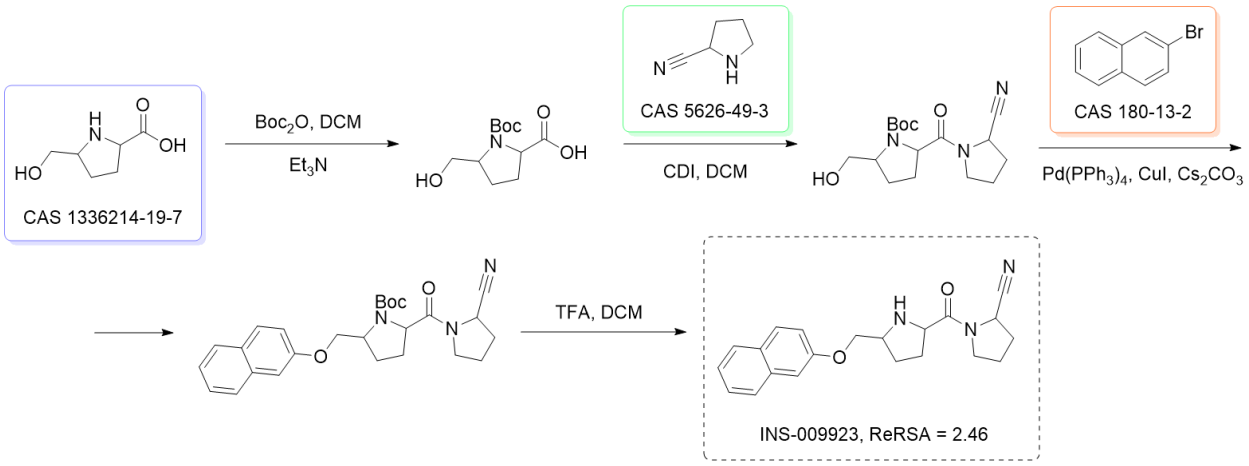


Схема 2. Предполагаемая синтетическая схема для соединения **INS-009923** из модельного эксперимента №1.

Таким образом, удалось легко восстановить синтетическую схему для избранного соединения **INS-009923** из модельного эксперимента 1, пользуясь результатами работы модуля ReRSA.

3.2.3.2 Валидация модуля ReRSA на основе модельного эксперимента № 2

Результат применения метода моделирования ReRSA к избранной молекулярной структуре **X1975-5319-0159** из модельного эксперимента № 2 (см. 3.1.3.2) представляет собой набор из 4 КДИС (см. рис. 54).

Синтетический анализ соединения **X1975-5319-0159** (см. схему 3), исходя из КДИС, предложенных методом ReRSA, может выглядеть следующим образом. Производное пиразина **CAS 28795-50-8** можно ввести в реакцию с альдегидом **CAS 50675-18-8**. При этом вводить защитную группу для второй вторичной аминогруппы не требуется, поскольку стерическое затруднение производимое карбинольным заместителем в подобного рода субстратах обеспечивает должный уровень региоселективности, как было показано, например, в синтезах фенебрутиниба (fenebrutinib [168]) или энсартиниба (ensartinib [169]). Полученный аддукт восстановительного аминирования затем можно последовательно ввести в реакцию арилирования по Бухвальду-Хартвигу с арилбромидом **CAS 1781878-14-5** и метилмеркаптаном **CAS 74-93-1**, что в итоге даст искомое соединение **X1975-5319-0159**.

X1975-5319-0159, ReRSA = 2.47

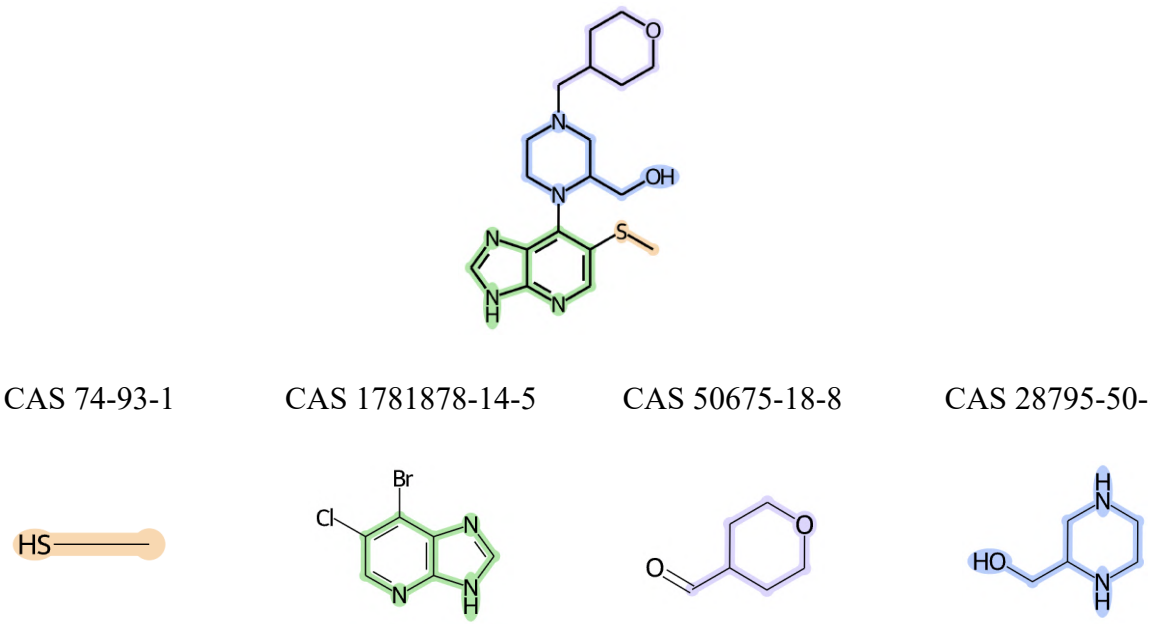


Рисунок 54. Набор КДИС, предложенный алгоритмом ReRSA для молекулярной структуры **X1975-5319-0159** из модельного эксперимента №2.

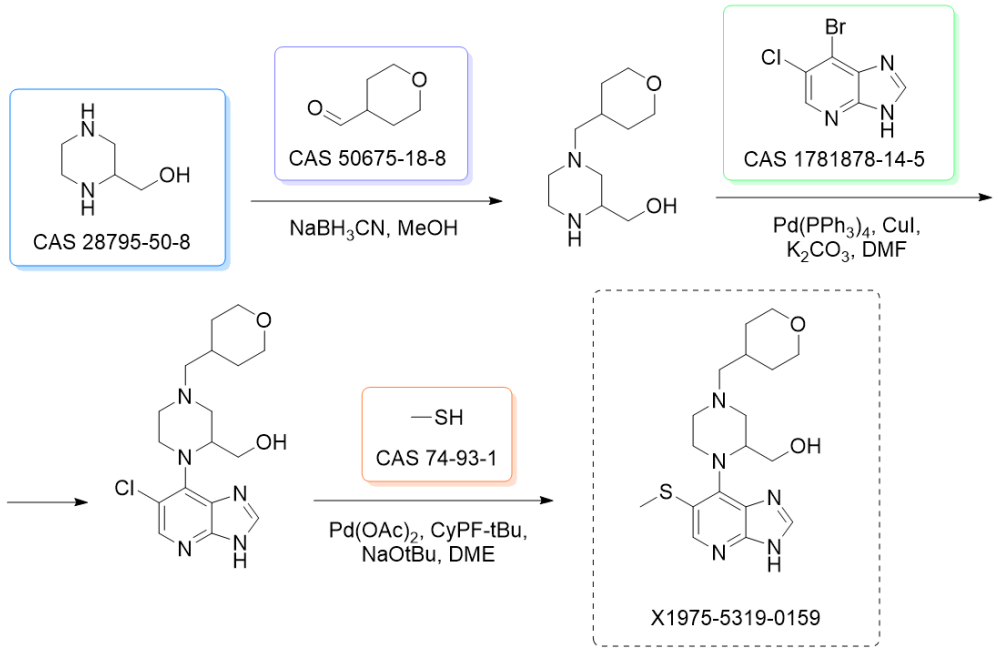


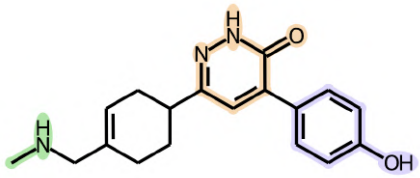
Схема 3. Предполагаемая синтетическая схема для молекулярной структуры **X1975-5319-0159** из модельного эксперимента №2.

Таким образом, удалось легко восстановить синтетическую схему для избранного соединения **X1975-5319-0159** из модельного эксперимента 2, пользуясь результатами работы модуля ReRSA.

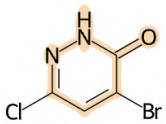
3.2.3.3 Валидация модуля ReRSA на основе модельного эксперимента № 3

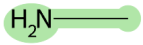
Результат применения метода моделирования ReRSA к избранной молекулярной структуре **X1975-8222-0001** из модельного эксперимента №3 (см. 3.1.3.3) представляет собой набор из 3 КДИС. Причем видно, что часть молекулярной структуры, содержащая циклогексеновый фрагмент, не была сконвертирована в КДИС, что, в том числе, нашло отражение в более высоком значении функции ReRSA, которое для молекулярных структур с полной конверсией в КДИС обычно находится в диапазоне < 3 (см. рис. 55).

X1975-8222-0001, ReRSA = 3.77



CAS 933041-13-5 CAS 74-89-5 CAS 269409-70-3





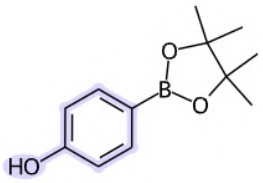


Рисунок 55. Набор КДИС, предложенный алгоритмом ReRSA для молекулярной структуры **X1975-8222-0001** из модельного эксперимента №3.

Синтетический анализ соединения **X1975-8222-0001** (см. схему 4), исходя из КДИС, предложенных методом ReRSA, может выглядеть следующим образом. Синтез соединения можно начать с двух последовательных реакций кросс-сочетания по Сузуки, сначала между найденными алгоритмом ReRSA исходных соединений борпинаколата **CAS 933041-13-5** и арилбромидом **CAS 933041-13-5**, затем между полученным на первой стадии продуктом реакции с борпинаколатом **CAS 1106871-37-7**, который был найден в базе данных КДИС SciFinder. Отметим, что данное КДИС есть и в базе данных, используемой в методе ReRSA, однако дальнейшие реакции из синтетического анализа, очевидно, не были заложены в метод ReRSA. Удаление на следующей стадии ацетальной защиты приводит к образованию кетона, который, однако, не может быть введен напрямую с литиевым производным дитиана без защиты фенольного гидроксила, для чего на следующей стадии можно провести классическую защиту фенольного гидроксила метильной группой. Присоединение литийорганического соединения на основе дитиана преследует обойти классическое ретросинтетическое

альтернирование зарядов, используя *Umpolung* стратегию синтеза. Полученный аддукт затем может быть обработан в кислых условиях, как для удаления дитиановой защиты, так и для дегидратации получаемого третичного спирта. Образующийся непредельный альдегид можно ввести в реакцию восстановительного аминирования с метиламином **CAS 74-89-5**, чтобы на последней стадии путем удаления метильной защитной группы с фенола классическим способом получить искомое соединение **X1975-8222-0001**.

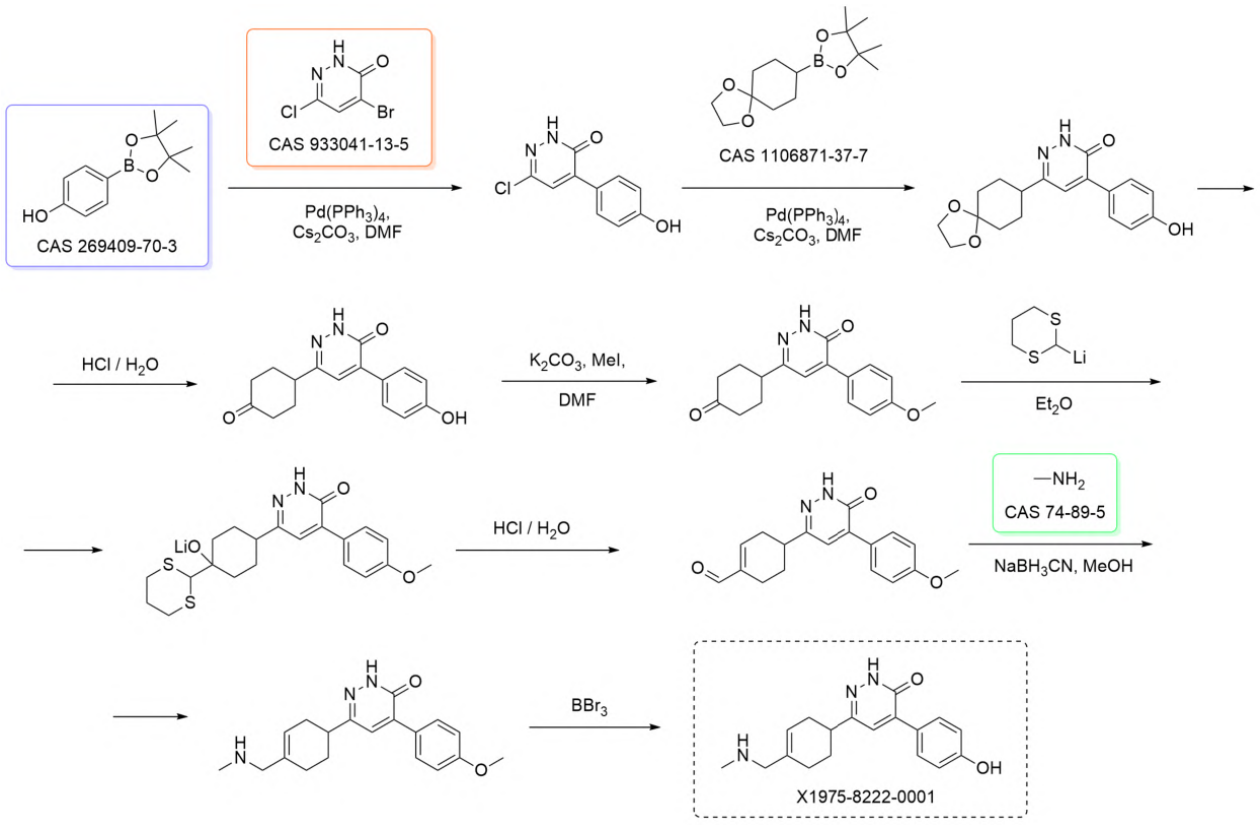


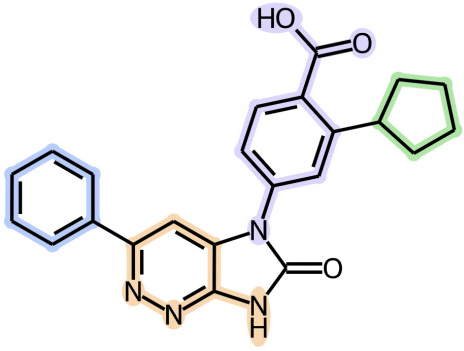
Схема 4. Предполагаемая синтетическая схема для молекулярной структуры **X1975-8222-0001** из модельного эксперимента №3.

Таким образом, удалось восстановить синтетическую схему для избранного соединения **X1975-8222-0001** из модельного эксперимента 3, пользуясь результатами работы модуля ReRSA. Стоит отметить, что поскольку модуль не представил информацию о конвертации всех частей молекулярной структуры в КДИС, потребовалось обращение к базе данных КДИС, а также дополнительные интеллектуальные усилия для того, чтобы продумать схему синтеза соединения, которая учитывала бы стратегию с использованием защитных групп и *Umpolung* стратегию, которая не покрывается методом ReRSA. Полученная схема синтеза состоит из 8 стадий, что отражается и на относительно высоком значении (3.77) функции ReRSA, в сравнении с другими молекулярными структурами, рассматриваемыми в этом разделе.

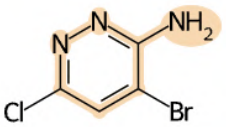
3.2.3.4 Валидация модуля ReRSA на основе модельного эксперимента № 4

Результат применения метода моделирования ReRSA к избранной молекулярной структуре **X1975-8261-0001** из модельного эксперимента №4 (см. 3.1.3.4) представляет собой набор из 4 КДИС (см. рис. 56).

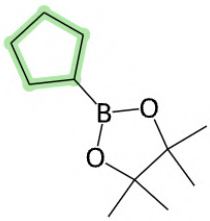
X1975-8261-0001, ReRSA = 2.43



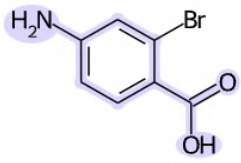
CAS 446273-59-2



CAS 66217-55-8



CAS 2486-52-4



CAS 98-80-6

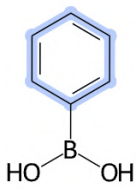


Рисунок 56. Набор КДИС, предложенный алгоритмом ReRSA для молекулярной структуры **X1975-8261-0001** из модельного эксперимента №4.

Синтетический анализ соединения **X1975-8261-0001** (см. схему 5), исходя из КДИС, предложенных методом ReRSA, может выглядеть следующим образом. Синтез соединения можно начать с предварительной защиты аминогруппы ароматической карбоновой кислоты **CAS 2486-52-4** во избежание ее арилирования на последующей стадии по реакции Чана-Лама. Защищенную Boc-защитной группой кислоту затем можно ввести в реакцию Сузуки с циклопентилборпинаколатом **CAS 66217-55-8.** Полученное производное о-циклопентил-п-аминокарбоновой кислоты имеет смысл напрямую вводить в реакцию фосгенирования с

ароматическим амином **CAS 446273-59-2**, поскольку образующийся в результате фосгенирования хлороводород может удалять Boc-защитную группу. Полученный аддукт реакции фосгенирования, который в ретро-реакциях ReRSA фрагментировался бы реакцией R27 (см. табл. 3), затем можно подвергнуть внутримолекулярной циклизации, которая может пройти без катализатора по механизму ароматического нуклеофильного замещения через присоединение-отщепление из-за высокой электрофильностью атома брома, активированного ароматическим атомом азота пиридазинового кольца. Дополнительным фактором, способствующим данной реакции, является, собственно, сама природа внутримолекулярной циклизации с положительным энтропийным эффектом. После реакции циклизации полученное вещество имеет смысл защитить по атому азота со свободной валентностью, прежде чем вводить фенилбороновую кислоту **CAS 98-80-6**, поскольку последняя в противном случае может вступить в конкурирующую реакцию Чана-Лама. После заключительного кросс-сочетания по Сузуки с **CAS 98-80-6** необходимо будет удалить Boc-защитной группу, что приведет к получению искомого вещества **X1975-8261-0001**.

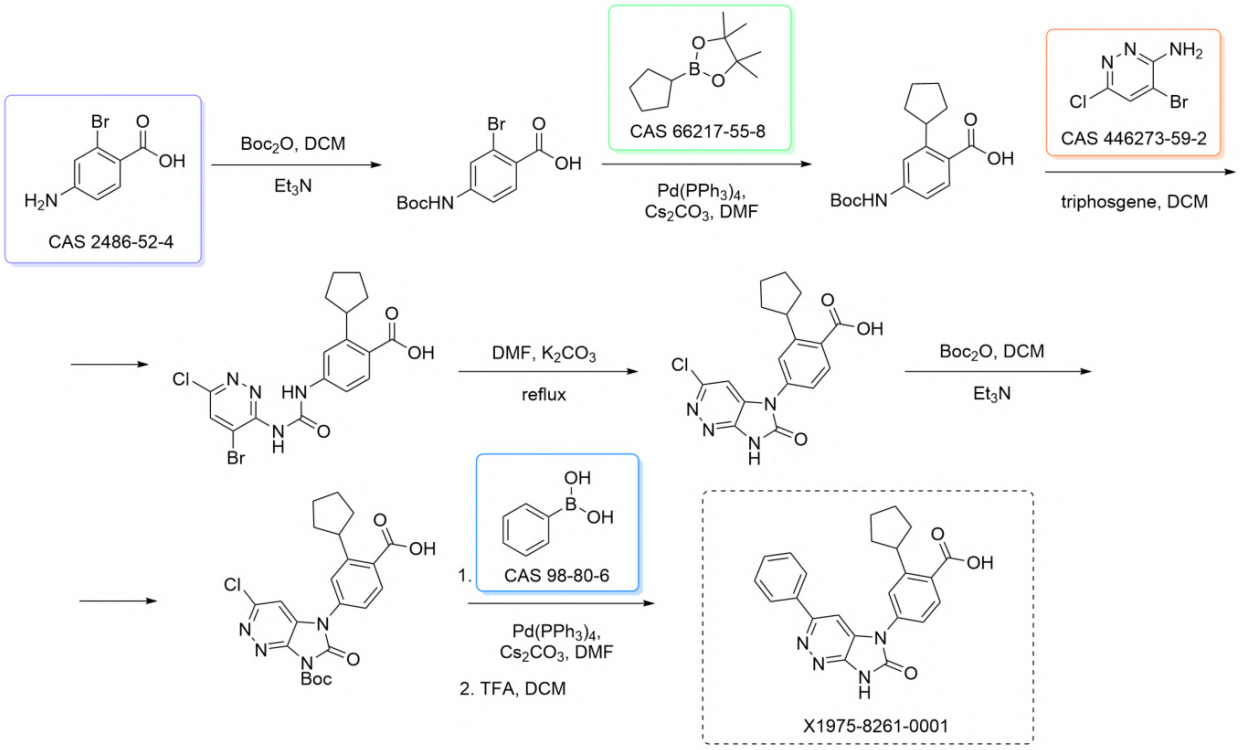


Схема 5. Предполагаемая синтетическая схема для молекулярной структуры **X1975-8261-0001** из модельного эксперимента №4.

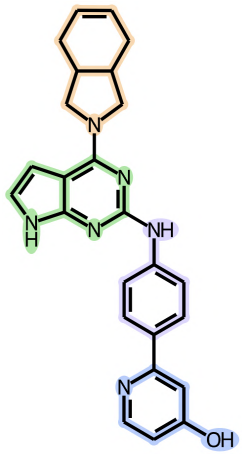
Таким образом, удалось легко восстановить 6–7-стадийную синтетическую схему для избранного соединения **X1975-8261-0001** из модельного эксперимента №4, пользуясь результатами работы модуля ReRSA и применяя знания о защитных группах и особенностях

региоселективности реакций с бороновыми кислотами и борпинаколатами, которые находятся за рамками метода ReRSA.

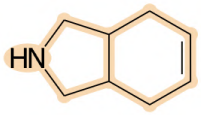
3.2.3.5 Валидация модуля ReRSA на основе модельного эксперимента № 5

Результат применения метода моделирования ReRSA к избранной молекулярной структуре **X1975-8230-0001** из модельного эксперимента №5 (см. 3.1.3.5) представляет собой набор из 4 КДИС (см. рис. 57).

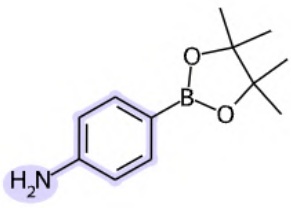
X1975-8230-0001, ReRSA = 2.67



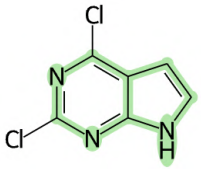
CAS 2144-87-8



CAS 214360-73-3



CAS 90213-66-4



CAS 36953-40-9

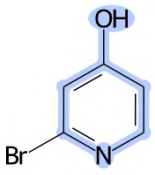


Рисунок 57. Набор КДИС, предложенный алгоритмом ReRSA для молекулярной структуры **X1975-8230-0001** из модельного эксперимента №5.

Синтетический анализ соединения **X1975-8230-0001** (см. схему 6), исходя из КДИС, предложенных методом ReRSA, может выглядеть следующим образом. Синтез соединения можно осуществить при помощи двух последовательных реакций ароматического

нуклеофильного замещения по механизму присоединение-отщепление. Согласно наблюдаемой региоселективности для производных 2,4-дихлорпиримидинов сначала будет происходить замещение атома хлора в положении 4, а затем в положении 2. Таким образом первую реакцию нужно проводить между вторичным амином **CAS 2144-87-8** и производным 2,4-дихлорпиримидина **CAS 90213-66-4**. Конвергентный синтез **X1975-8230-0001** имеет смысл завершить подобной реакцией по положению 2 с продуктом параллельной ветви синтеза, который, в свою очередь, может быть получен по реакции кросс-сочетания по Сузуки между 2-бром,4-гидроксипиридином **CAS 36953-40-9** и борпинаколатом **CAS 214360-73-3.**

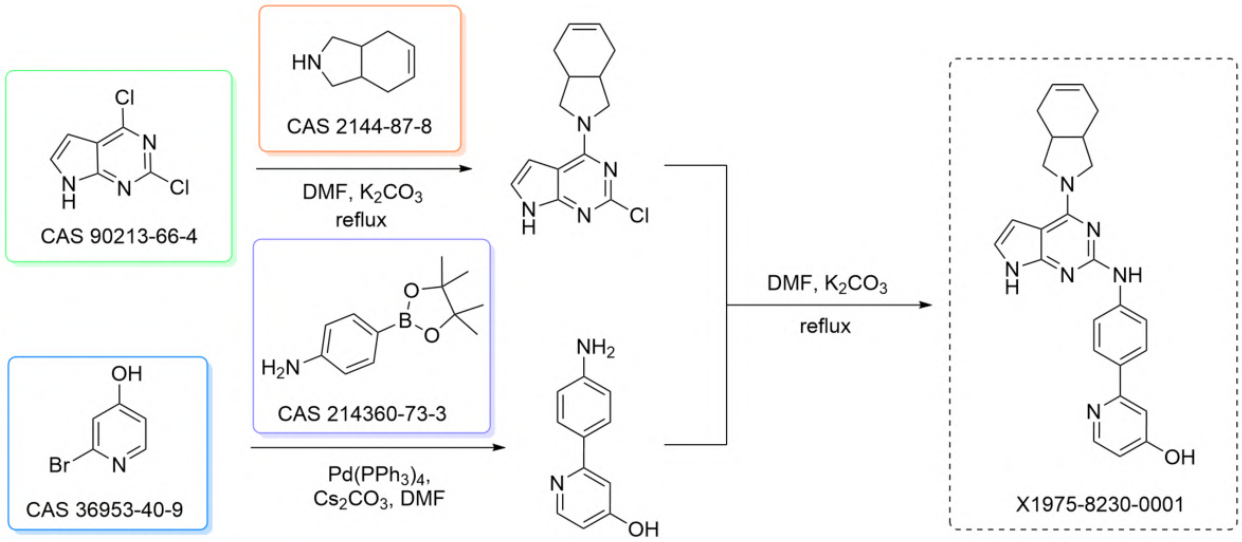


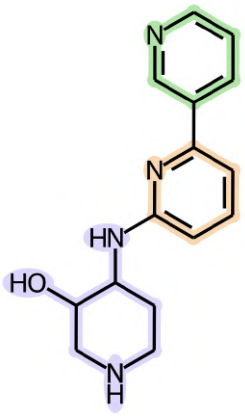
Схема 6. Предполагаемая синтетическая схема для молекулярной структуры **X1975-8230-0001** из модельного эксперимента №5.

Таким образом, удалось легко восстановить 3-стадийную конвергентную синтетическую схему для избранного соединения **X1975-8230-0001** из модельного эксперимента №5, пользуясь результатами работы модуля ReRSA.

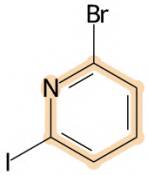
3.2.3.6 Валидация модуля ReRSA на основе модельного эксперимента № 6

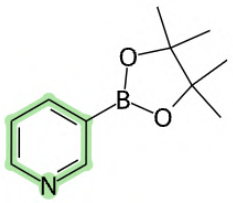
Результат применения метода моделирования ReRSA к избранной молекулярной структуре **X1975-4541-0002** из модельного эксперимента №6 (см. 3.1.3.6) представляет собой набор из 3 КДИС (см. рис. 58).

X1975-4541-0002, ReRSA = 2.19



CAS 234111-08-1 CAS 329214-79-1 CAS 1388089-28-8





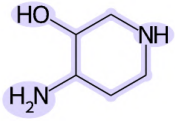


Рисунок 58. Набор КДИС, предложенный алгоритмом ReRSA для молекулярной структуры **X1975-4541-0002** из модельного эксперимента №6.

Синтетический анализ соединения **X1975-4541-0002** (см. схему 7), исходя из КДИС, предложенных методом ReRSA, может выглядеть следующим образом. Синтез соединения можно осуществить при помощи двух последовательных реакций кросс-сочетания: в начале провести реакцию Сузуки между борпинаколатом **CAS 329214-79-1** и 2-бром,6-йодпиридином **CAS 234111-08-1**, а затем провести реакцию Бухвальда-Хартвига с амином **CAS 1388089-28-8**. Несмотря на возможные вопросы к региоселективности подобной реакции, ряд опубликованных ранее результатов говорит о том, что можно селективно провести N-арилирование первичной аминогруппы в присутствии вторичной аминогруппы в подобных гетероалифатических диаминах [170–172].

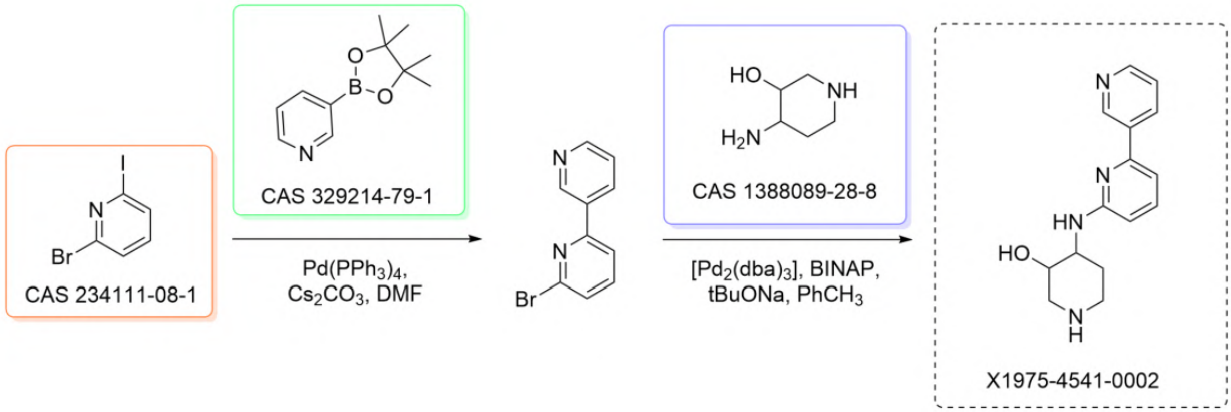


Схема 7. Предполагаемая синтетическая схема для молекулярной структуры **X1975-4541-0002** из модельного эксперимента №6.

Таким образом, удалось легко восстановить 2-стадийную схему для избранного соединения **X1975-4541-0002** из модельного эксперимента №6, пользуясь результатами работы модуля ReRSA.

Если суммировать результаты синтетического анализа потенциальных лекарственных веществ, произведенных платформой генеративной химии Chemistry42 в ходе выполнения модельных экспериментов 1-6, то можно сделать однозначный вывод о том, что метод моделирования синтетической доступности ReRSA предоставляет полезную информацию о КДИС для наилучших (с точки зрения оценочных модулей платформы) молекулярных структур, используя которую медицинский химик может восстановить синтетическую схему, зачастую даже не обращаясь к референсным базам данным КДИС и химических реакций, таких как Reaxys или SciFinder. Сам факт того, что модуль ReRSA способен предоставлять информацию подобного рода и качества говорит о том, что пользователь может с высокой степенью достоверности опираться на значения ReRSA в целях оценки синтезируемости молекулярных структур, получаемых при помощи платформы Chemistry42.

3.2.3.7 Валидация модуля ReRSA на предмет фильтрации синтетически нерелевантных 5-членных ароматических гетероциклов среди результатов платформы генеративной химии

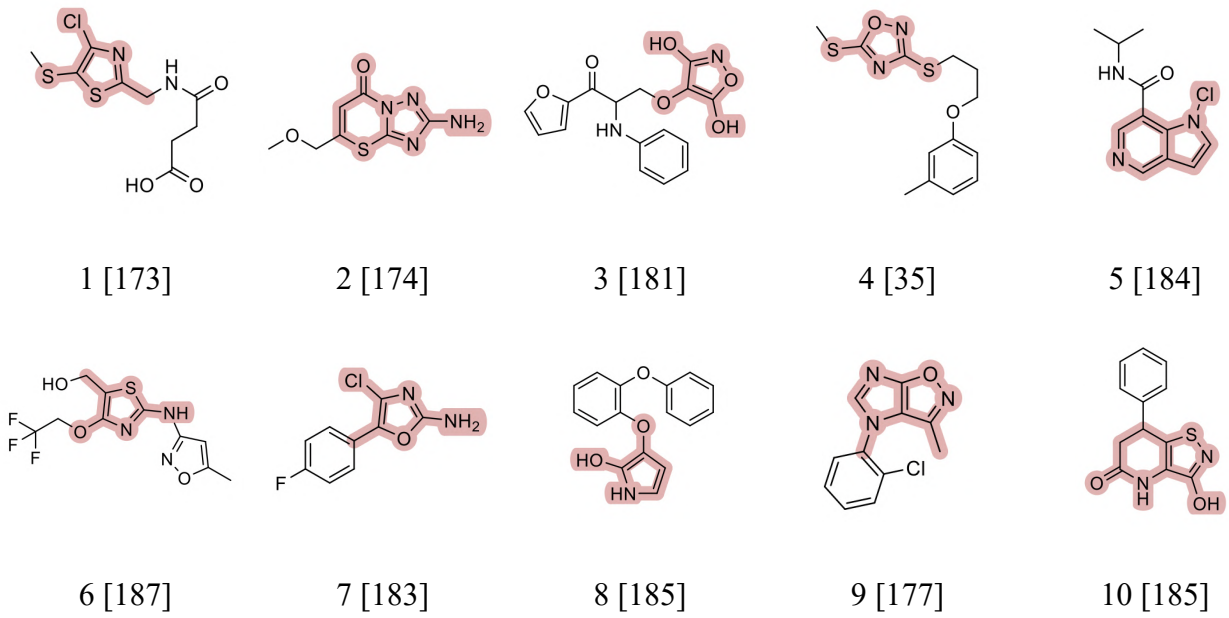
Отдельно была валидирована часть алгоритма ReRSA, ответственная за штрафование молекулярных структур, содержащих синтетически нерелевантные подструктуры. Был собран датасет молекулярных структур, сгенерированными современными генеративными моделями. Доля отфильтрованных молекулярных структур варьируется от 0.3% до 28.5% (см. табл. 35).

Таблица 35. Результаты анализа выходных данных современных генеративных моделей.

№ набора	Число структур				Ссылка
	в опубликованном наборе	невалидных*	отфильтровано		
1	30 000	0	1 933	(6.44%)	[35]
2	454 700	619	6 614	(1.46%)	[173]
3	2 106 904	0	4 202	(0.20%)	[174]
4	220 377	1	14 433	(6.55%)	[175]
5	1 267	0	21	(1.66%)	[176]
6	6 106	0	158	(2.59%)	[177]
7	3 183	0	50	(1.57%)	[178]
8	188	0	9	(4.79%)	[179]
9	7 052	0	68	(0.96%)	[180]
10	200	0	57	(28.50%)	[181]
11	381 579	32 516	4 386	(1.26%)	[182]
12	20 000	1	447	(2.24%)	[183]
13	15 000	42	258	(1.72%)	[184]
14	7 080	0	21	(0.30%)	[185]
15	12 060 315	0	559 942	(4.64%)	[186]
16	4 041	0	50	(1.24%)	[187]

* Невалидными считались структуры, не прошедшие стандартную процедуру проверки и нормализации RDKit. Отметим, что молекулярные структуры оценивались только с точки зрения валидности их SMILES-строки и синтетической доступности в контексте пятичленных ароматических гетероциклов, не оценивалось соответствие структур drug-like критериям.

Примеры отфильтрованных молекулярных структур, содержащих синтетически нерелевантный фрагмент, приведены на рис. 59. Следует отметить, что молекулярные структуры оценивались только с точки зрения валидности их SMILES-строки и синтетической доступности в контексте пятичленных ароматических гетероциклов, и не оценивалось соответствие структур *drug-like* критериям.



1 [173] 2 [174] 3 [181] 4 [35] 5 [184]

6 [187] 7 [183] 8 [185] 9 [177] 10 [185]

Рисунок 59. Сгенерированные молекулярные структуры, содержащие синтетически нерелевантные фрагменты.

Помимо указанного эксперимента, были проанализированы выходные данные модельных экспериментов Chemistry42 с ReRSA 2.0. Показано, что доля не прошедших фильтр структур колеблется от 3 до 39% (табл. 36).

Таблица 36. Результаты анализа выходных данных модельных генераций платформы Chemistry42

Название модельного эксперимента	**Молекулярных структур**	
	сгенерировано	**отфильтровано**
De novo дизайн ингибиторов Jak3 киназы	4 419	1 296 (29%)
Дизайн аналогов соединения-хита протеазы USP7	4 378	425 (10%)
Генеративный *de novo* дизайн ингибиторов тирозинкиназы рецептора эпидермального фактора роста (EGFR)	2 812	905 (32%)
Генеративный *scaffold-hopping* дизайн ингибиторов CAMKK2 киназы	1 512	90 (6%)
Дизайн заместителей ингибитора MPS1 киназы	7 334	252 (3%)
Дизайн ингибиторов главной протеазы коронавируса SARS-CoV-2 на основе знаний о связывании малого фрагмента	1 968	120 (6%)

Также были проанализированы генерации Chemistry42 с зафиксированным пятичленным ароматическим фрагментом с помощью функционала якорных точек (табл. 37). Указанные условия обязывают каждую сгенерированную молекулярную структуру содержать пятичленный ароматический гетероцикл. Показано, что в таком случае доля отсеянных разработанных алгоритмом структур возрастает и составляет от 28% до 39%.

Таблица 37. Результаты анализа выходных данных генераций платформы Chemistry42 с зафиксированным пятичленным ароматическим циклом

Название модельного эксперимента	**Молекулярных структур**	
	сгенерировано	**отфильтровано**
Дизайн ингибиторов фосфоинозитид-3-киназы (PI3K) — аналогов алпелизиба, на основе структуры мишени	5 102	1 994 (39%)
Дизайн ингибиторов тирозин-протеинкиназы ABL2 — аналогов дазатиниба, на основе структуры мишени	2 566	706 (28%)
Дизайн ингибиторов тирозин-протеинкиназы RET — аналогов селперкатиниба, на основе структуры мишени	3 992	1 464 (37%)

Продемонстрированные результаты *in silico* экспериментов показывают эффективность в отсеве синтетически нерелевантных структур и подчеркивают актуальность разработанного нами алгоритма фильтрации.

3.2.3.8 In silico валидация модуля ReRSA на основе результатов моделирования синтетической доступности ретросинтетическим модулем платформы Chemistry42

В настоящий момент нашей командой параллельно ведется разработка полноценного движка автоматизированного ретросинтеза молекулярных структур (CASP-система), в основу которого лег запатентованный нами алгоритм [75]. В сентябре 2024 г. была выпущена альфа-версия продукта для внутреннего тестирования и пользования, а в декабре того же года стартовал открытый бета-тест ретросинтетического модуля платформы Chemistry42. На текущий момент пользователям платформы доступна версия 1.0 ретросинтетического модуля, которая позволяет в автоматическом режиме проводить ретросинтетический анализ до 1 000 молекулярных структур в одном эксперименте (2–8 часов на 1 000 структур в зависимости от их структурной сложности). И хотя настоящее исследование не ставит цель описать ретросинтетический модуль и вклад автора диссертации в его разработку, представляется

показательным эксперимент по валидации модуля ReRSA с помощью вышеупомянутого CASP-модуля.

Суть валидационного эксперимента заключается в сравнении результатов моделирования синтетической доступности новых молекулярных структур алгоритмом ReRSA и ретросинтезом, как наиболее надежным способом моделирования СД. И хотя в разд. 1.4.3.1 упоминались недостатки CASP-систем, из подобного сравнения двух независимых ММСД, основанных на разных подходах, можно извлечь ценные выводы, тем более что арсенал платформы Chemistry42 и возможности ретросинтетического модуля, в частности, позволяют провести масштабное исследование в разумные сроки.

Объектом эксперимента являются новые молекулярные структуры, сгенерированные на платформе Chemistry42. Источниками послужили 5 генеративных модельных экспериментов (№2–6, см. разд. 3.1.3, табл. 13), а так же 9 SBDD генеративных экспериментов с клинически релевантными новыми мишенями [188–191] в качестве входных данных (см. табл. 38). Во всех указанных генеративных экспериментах модуль ReRSA был отключен, потому генеративные модели не были ограничены с точки зрения синтетической доступности генерируемых молекулярных структур.

Всего было сгенерировано 151 385 уникальных структур. Для всех молекулярных структур были рассчитаны значения ReRSA в “жесткой политике” (см. разд. 2.2.10), распределение молекулярных структур по диапазонам значений ReRSA приведено на рис. 60.

Таблица 38. Источники и количества молекулярных структур для валидации модуля ReRSA при помощи автоматизированного ретросинтеческого движка (для 9 клинически релевантных новых мишеней указан код PDB, выбранного для SBDD генерации)

#	ID эксперимента	Число сгенерированных молекулярных структур
Модельные эксперименты		
1	№2	25 969
2	№3	4 396
3	№4	3 358
4	№5	6 132
5	№6	11 555
SBDD генерации		
6	3W2R	33 735
7	5WG6	14 616
8	5YVE	7 576
9	6VKV	7 361
10	6FS8	5 536
11	8ONV	5 186
12	8DZ0	4 923
13	7W80	4 076
14	5ZR3	16 966
Всего уникальных структур		**151 385**

Отключение модуля ReRSA позволило набрать репрезентативную выборку молекулярных структур относительно каждого диапазона из возможных значения ReRSA.

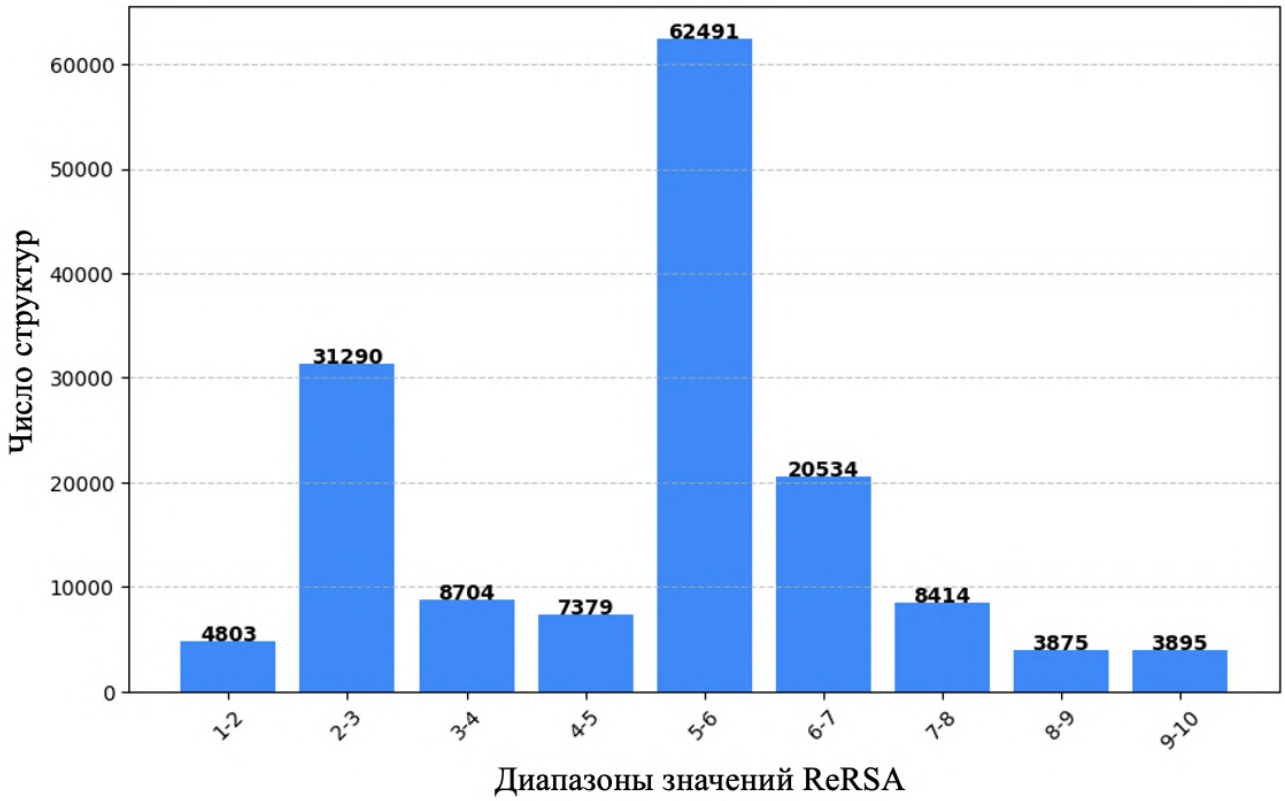


Рисунок 60. Распределение сгенерированных молекулярных структур по диапазонам значений ReRSA.

Для ретросинтетического анализа случайным образом было отобрано по 1 000 молекулярных структур, соответствующих каждому из диапазонов значений ReRSA [1–2), [2–3), …, [9–10], с сохранением репрезентативности молекулярных структур из разных генеративных экспериментов-источников (см. рис. 61). Отобранные молекулярные структуры подвергались автоматическому анализу ретросинтетическим модулем со следующими настройками: (1) до 15 минут на поиск ретросинтетического пути на одну структуру; (2) максимальная длина пути — 10 шагов; (3) разрешено образование рацематных смесей, подразумевающее разделение энантиомеров на последнем шаге; (4) использован набор КДИС, используемый для расчета значений и визуализации ReRSA.

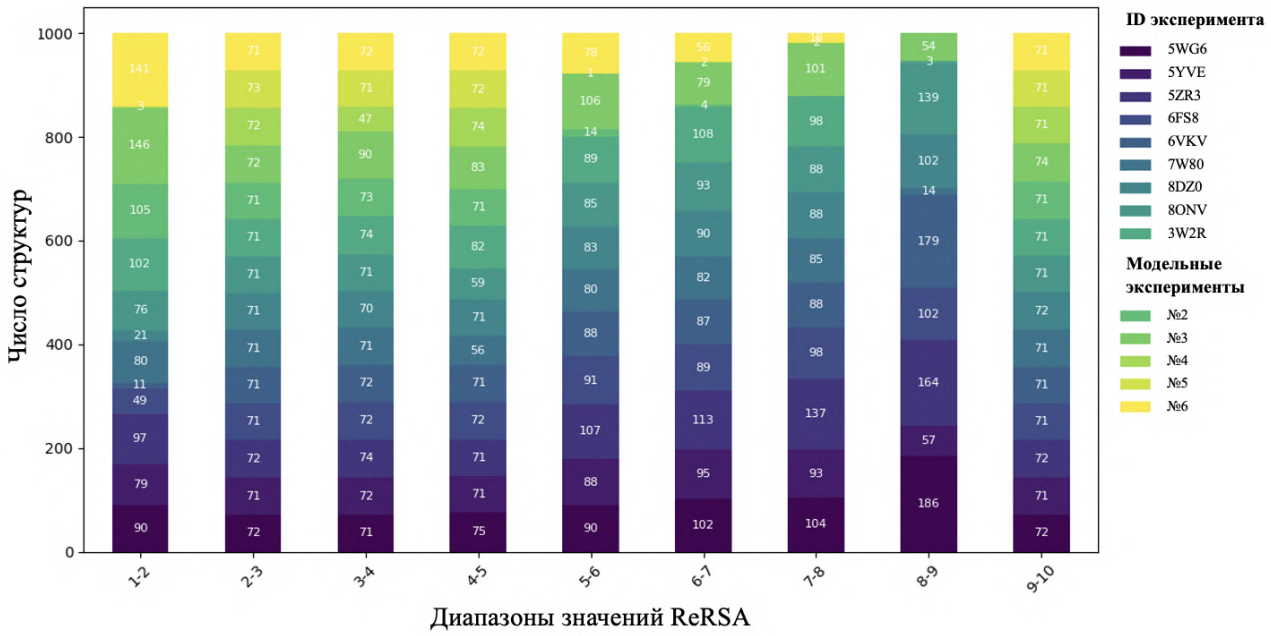


Рисунок 61. Распределение отобранных молекулярных структур по источникам и диапазонам значений ReRSA.

Доля успешно решенных соединений (*ретросинтетическая сходимость*, доля соединений, для которых ретросинтетический модуль нашел хотя бы 1 путь, сходящийся к КДИС в рамках перечисленных выше ограничений поиска) для каждого из диапазонов ReRSA приведена на рис. 62. Диаграмма демонстрирует сильную корреляцию между оценкой ReRSA и ретросинтетической сходимостью. Ретросинтетическая сходимость снижается с ростом значений ReRSA. Более 90% структур со значением ReRSA от 1 до 2 сошлись к ретросинтетическому пути, в то время как для структур со значением ReRSA от 5 до 8 сошлись лишь около 20% структур. В конечном диапазоне оценок ReRSA (9–10) сходимость, на первый взгляд, аномально высокая. Однако, это объясняется встроенным в модуль ReRSA фильтром синтетически нерелевантных подструктур (см. разд. 2.2.9 и 3.2.3.7). Алгоритм ReRSA присваивает оценку 10 молекулярным структурам, содержащим фрагменты (замещенные пятичленные ароматические гетероциклы), не встречающиеся в известном химическом пространстве, что косвенно может свидетельствовать о сложности или вовсе невозможности их синтеза. Тем не менее, ретросинтетические алгоритмы всё ещё способны генерировать правдоподобные синтетические пути для таких структур, хотя их практическая осуществимость остаётся неопределённой из-за отсутствия прецедентов.

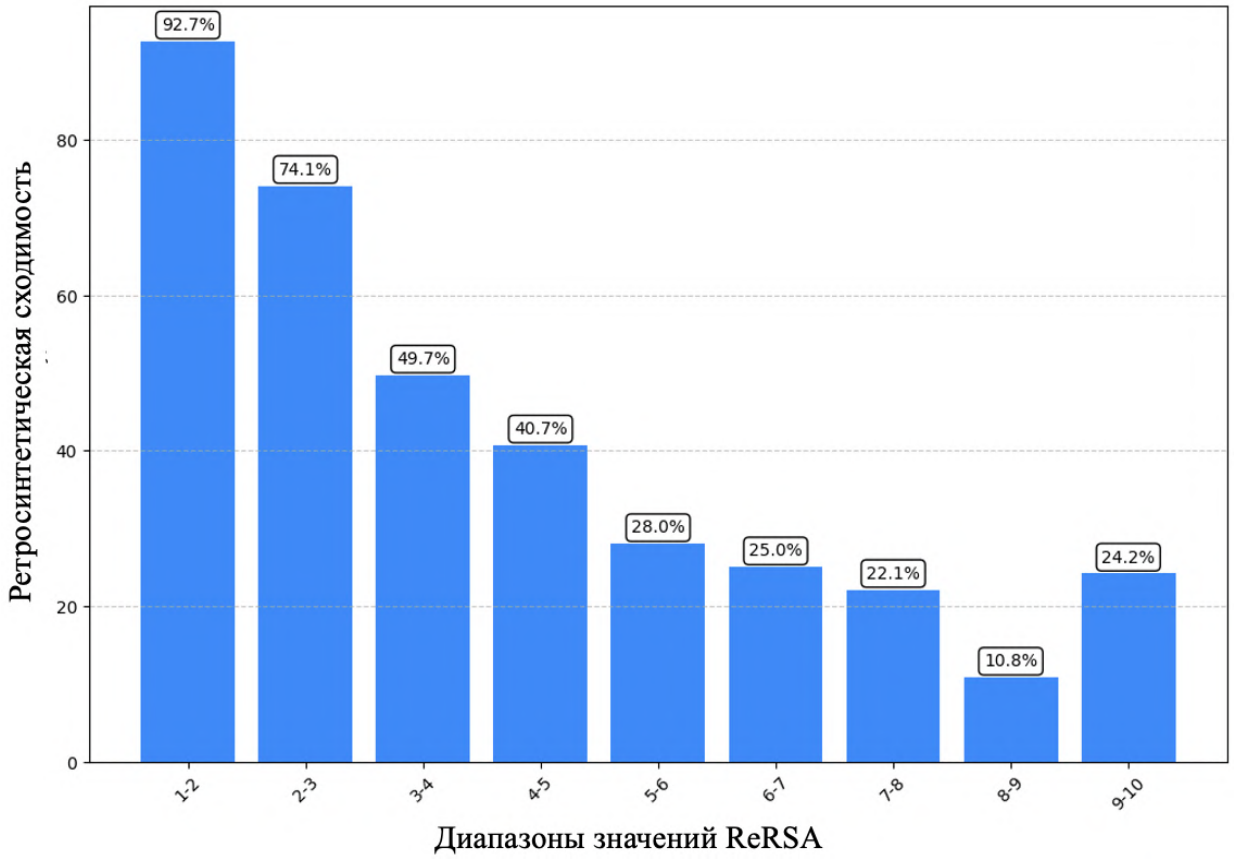


Рисунок 62. Ретросинтетическая сходимость молекулярных структур в зависимости от значений ReRSA.

Настоящий эксперимент подтверждает гипотезу о том, что, оценка ReRSA коррелирует с синтетической сложностью молекулярной структуры, даже несмотря на то, что “химический арсенал” модуля ReRSA, содержащий всего 52 вида реакций органического синтеза (см. п. 2.2.4), заметно уступает таковому автоматизированного ретросинтетического движка, который содержит около 3000 реакций. И не смотря на то, что для более высоких значений ReRSA (диапазон [7–10]) ретросинтетическая сходимость молекулярных структур ненулевая, утверждать, что эти сошедшиеся структуры, имеющие высокое значение ReRSA, являются ложно отрицательными срабатываниями ММСД, нельзя, поскольку ретросинтез хоть и является “золотым стандартом” моделирования синтетической доступности, тем не менее не является гарантом принципиальной синтезируемости молекулярной структуры, что может быть подтверждено только успешным синтезом в лаборатории. Точно по той же причине нельзя утверждать и обратное — если структура не сошлась ретросинтетическим образом при помощи CASP-инструмента, это вовсе не означает, что она сложно синтезируемая или неосуществима вовсе. Тем не менее, эксперимент наглядно демонстрирует корреляцию прогнозов двух независимых ММСД — ретросинтетического модуля платформы Chemistry42 и модуля ReRSA.

3.2.4 Производительность алгоритма ReRSA

Поскольку скорость работы модуля является чрезвычайно важным факторов для обработки больших потоков молекулярных структур, производимых генеративными моделями, контроль среднего времени обработки одной молекулярной структуры выполнялся на каждом этапе модификации алгоритма. Итоговые времена, актуальные для версии 3.0, представлены в таблице 39.

Таблица 39. Производительность модуля ReRSA, измеренная во время модельных экспериментов

Модельный эксперимент	**Среднее время обработки одной структуры модулем ReRSA, с**
№2	0.0312
№3	0.0410
№4	0.0564
№5	0.0844
№6	0.0295

Не смотря на то, что среднее время обработки молекулярных структур варьируется между модельными экспериментами в силу наличия разных хемотипов, показано, что среднее время обработки модулем ReRSA одной структуры не превышает 0.1 секунды, что, с одной стороны, на порядок дольше, чем среднее время обработки одной структуры алгоритмом SA Score (в оригинальной статье производительность оценивается на уровне 100 000 молекулярных структур за 3 минуты, что равно 0.0018 секунд на одну структуру [43]), но, тем не менее, значительно превосходит по скорости работы классические ретросинтетические движки, такие как Chematica или AiZynthFinder.

3.2.5 Недостатки алгоритма ReRSA

Несмотря на серьезные преимущества метода ReRSA над другими ММСД, что включает и использование ретросинтетической логики в результаты моделирования СД при сохранении высокой скорости вычислений (см. разд. 3.2.4), и учёт локальных структурных контекстов подструктур (см. разд. 2.2.9), а также возможность получать информацию о КДИС, ретросинтетически ассоциированных с целевой молекулой, у разработанного нами метода есть недостатки.

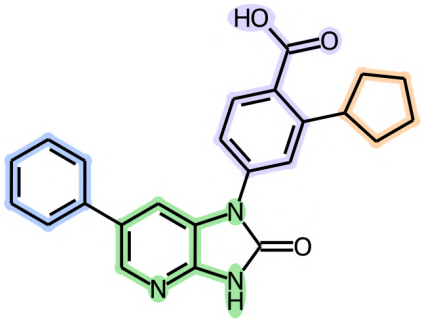
В качестве недостатков мы видим:

1. Отсутствие учёта стереохимии целевой молекулы и КДИС;
2. Невозможность учитывать логику региоселективности и хемоселективности;
3. Ограниченность по количеству робастных реакций, применяемых в методе.

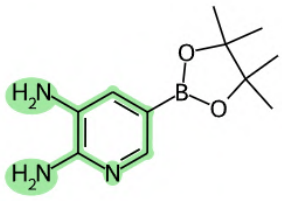
Наиболее серьезным недостатком мы видим именно невозможность учёта стереохимических факторов в ходе квази-ретросинтетической фрагментации. К сожалению, современные инструменты хемоинформатики, такие как RDKit не способны проводить реакции с учётом стереоспецифичности и стереоселективности. Поскольку методы BRICSDecompose и RunReactans, являющиеся базовыми для ММСД ReRSA, входят в состав библиотеки RDKit, то метод ReRSA наследует все проблемы, связанные с некорректным учётом стереохимии от RDKit. Промежуточным решением проблемы стало то, что пользователю предоставляется полный набор CAS идентификационных номеров для всех возможных стереоизомеров инвариантов КДИС. Безусловно, хоть это и усложняет работу на уровне пользовательского взаимодействия, но, тем не менее, позволяет пользователю найти из списка предлагаемых идентификационных номеров тот изомер КДИС, который соответствует требуемому изомеру целевой молекулы.

В целях минимизации проблем, связанных с региоселективностью и хемоселективностью, было предложено минимальное технически доступное решение, осуществленное в рамках метода: предоставление пользователю всех возможных вариантов по конвертации синтоноподобных фрагментов в синтетические эквиваленты, а затем в КДИС (см. рис. 63–64). Предполагается, что из предоставляемого набора гомологичных КДИС пользователь может выбрать тот, который подходит по соображениям региоселективности или хемоселективности.

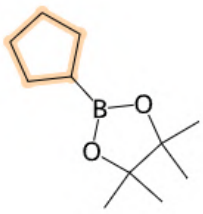
X1975-8261-0002, ReRSA = 2.43



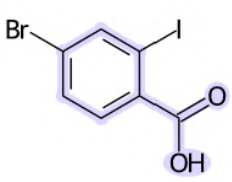
CAS 1204334-21-3



CAS 66217-55-8



CAS 1133123-02-0



CAS 108-90-7

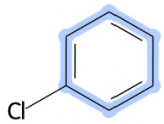


Рисунок 63. Набор КДИС, предложенный алгоритмом ReRSA для молекулярной структуры **X1975-8261-0002** из модельного эксперимента №4.

Например, в синтетической схеме молекулярной структуры **X1975-8261-0002** (см. схему 8) из модельного эксперимента №4 (см. табл. 21), которую пользователь может восстановить на основе предлагаемых КДИС, выбор **CAS 1133123-02-0** из гомологичного набора (см. рис. 59) соответствующих КДИС для данной части молекулы (сиреневая подсветка) может быть обусловлен как раз соображениями региоселективности.

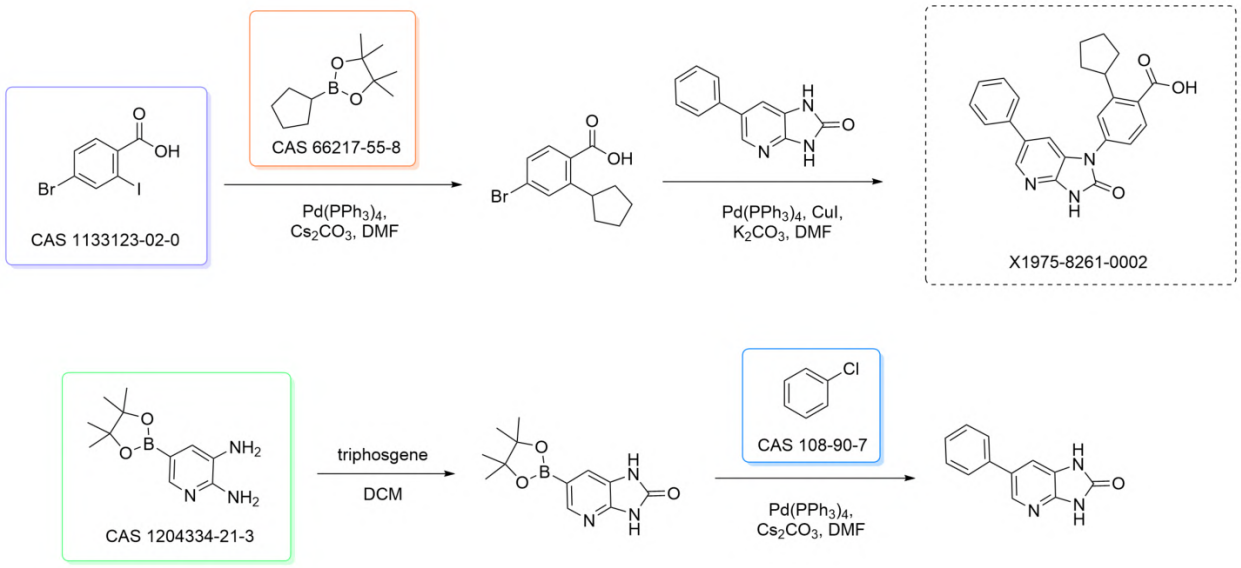
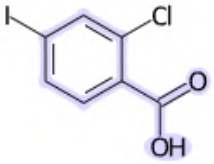


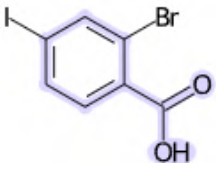
Схема 8. Предполагаемая схема синтеза для эталонного соединения **X1975-8261-0002** из модельного эксперимента №4.

Безусловно, в подобном наборе может и не найтись соединения, который соответствует критериям региоселективности или хемоселективности, что представляет риски использования метода.

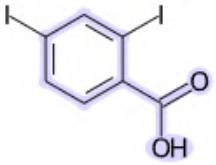
CAS 145343-76-6



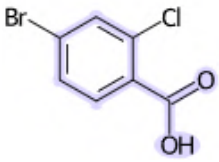
CAS 28547-29-7



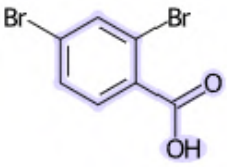
CAS 33522-84-8



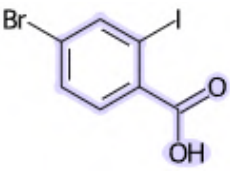
CAS 59748-92-4



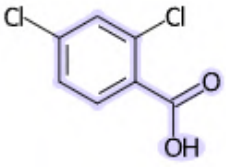
CAS 611-00-7



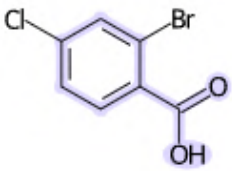
CAS 1133123-02-0



CAS 50-84-0



CAS 936-08-3



CAS 13421-13-1

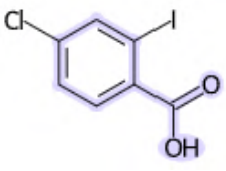


Рисунок 64. Предлагаемый пользователю набор КДИС, соответствующий одной из частей молекулярной структуры **X1975-8261-0002**.

Что касается хемоселективности и частично региоселективности, то техническое решение, предлагаемое в рамках метода, позволяет рассматривать КДИС, содержащие защитные группы. Предполагается, что пользователь сможет выбрать те исходные соединения, которые подходят по соображению селективности в рамках тех синтетических схем, которые пользователь может построить на основании КДИС. Но, как и в случае с региоселективностью, нахождение нужной защитной группы или комбинации защитных групп, не гарантировано методом.

Если говорить о том наборе робастных реакций, которые имплементированы в методе, то технических возможностей значительно расширять число реакций в наличии нет. По оценкам скорость работы модуля замедляется по гиперболической зависимости от числа робастных реакций. Особую чувствительность представляют основные “рабочие лошадки” метода — реакции, которые основаны на методе BRICSDecompose. Добавление даже одной подобной реакции может замедлить общую работу метода на 20–25%. Тем не менее, мы открыты для рассмотрения новых кандидатов в список робастных реакций по мере поступления обратной связи от пользователей платформы.

3.3 Метод моделирования структурных трендов медицинской химии MCE-18

3.3.1 Предварительные замечания об анализе структурных трендов в медицинской химии

Прежде чем приступить к основному обсуждению результатов по моделированию структурных трендов медицинской химии при помощи функции MCE-18, необходимо изложить некоторые концепции и результаты исследований, которые предшествовали тому, что представляет собой методология MCE-18, помимо тех, что были упомянуты в обзоре литературы.

В фундаментальном обзоре [114], посвященном развитию синтетических методологий, применяемых в медицинской химии, Н. Шнайдер и её коллеги помимо освещения основной темы обзора сделали попытку проследить структурные тренды на основании анализа структур продуктов реакций из патентной базы данных USPTO (англ., *United States Patent and Trademark Office*) с 1976 по 2015 год. Авторы обзора отметили увеличение общего числа насыщенных богатых sp^3-гибиридированными атомами углерода колец; однако они не представили детального количественного и качественного анализа структурной и пространственной сложности этих фрагментов. Авторы также исследовали эволюцию ключевых молекулярных свойств (молекулярная масса, LogP, PSA, доля sp^3-атомов, число

акцепторов и доноров водородных связей, а также доля свободно вращающихся связей) во времени. Они показали, что продукты реакций, раскрытые в фармацевтических записях, становились более крупными, более липофильными, более жёсткими и более растворимыми в течение исследуемого периода. Например, они отмечают, что средняя молекулярная масса и липофильность целевых молекул, заявленных в 1976 году, составляли 331 Да и 3,1 соответственно, тогда как в 2015 году эти значения увеличились в целом на 24% (MW = 409 Да) и 16% (LogP = 3,6). В период 1982–2010 гг. также увеличилось количество уникальных продуктов с молекулярной массой, выходящей за пределы правила пяти (MW > 500) — рост составил 16,5% (значение уклона 0,6), после чего наблюдалось небольшое снижение. Однако это уменьшение молекулярной массы можно отнести на счёт недостаточного объёма статистических данных к тому моменту (2015 год), поскольку статья группы Н. Шнайдер была опубликована в начале 2016 года. Площадь полярной поверхности молекулярных структур (PSA) увеличилась за соответствующий период более чем на 15% за исследуемый период, несмотря на одновременный рост липофильности, тогда как доля sp^3-гибридизированных атомов углерода, доля гетероатомов, а также количество акцепторов (HBA) и доноров (HBD) водородных связей не показали устойчивых тенденций. Авторы связали резкое снижение доли свободно вращающихся связей (FRB, *fraction of rotatable bonds*) и стагнацию (или небольшое снижение) показателя Fsp^3 с ростом жёсткости молекул. Однако, мы считаем, что это скорее следствие, чем причина, что будет обсуждено ниже. Более того, описанные выше результаты были получены с использованием необработанного набора данных; в частности, авторы не уменьшили число соединений с высокой структурной схожестью внутри перегруженных кластеров.

Мы считаем, что основная проблема, по которой авторы вышеупомянутого исследования сделали вывод о “стагнации” и “эволюционном спаде” заключается в том, что набор данных, который они использовали для анализа, не был должным образом проанализирован с точки зрения индустрии разработки лекарственных молекул и соответствующим образом подготовлен. Например, авторы обзора не задавались вопросом, привносят ли новизну молекулярные структуры, которые появляются в базах данных несколько раз в разные периоды времени по причинам не относящимся непосредственно к новым изобретениям, как, например, патенты на способ использования. Более того, отдельные гиперэкспрессированные молекулярные структуры можно расширить до целых гиперэкспресированных структурных кластеров, которые создают статистический шум при анализе химического пространства на предмет его развития. Особенно сильно могут влиять негативным образом молекулярные кластера, соответствующие природным соединениям,

таким как сапонины, алкалоиды, глюканы, олигосахариды, жирные кислоты, макролиды, катехины, морфины, подофиллотоксины, антрациклины, пенициллины и многие другие структурные классы. Все эти соединения и структуры со сходным строением существенно влияют на значение Fsp^3, что приводит к вводящим в заблуждение выводам и ложноположительным результатам. Таким образом, предварительно было необходимо очистить набор данных от гиперэкспресированных молекулярных структур и хемотипов и удостовериться в том, что записи, содержащие информацию о запатентованных ранее молекулярных структурах, не дублируются с хронологически более поздними записями. В таком случае, более поздние дублирующие записи должны были быть удалены, так как не имели отношения к развитию химического пространства.

3.3.2 Компоненты функции MCE-18 как дискриминирующие факторы при анализе структурных трендов медицинской химии

Предлагая вместо несколько наивного дескриптора Fsp^3 взвешенную долю sp^3-гибридизированных атомов углерода, находящихся в циклах (NCSPTR), мы показываем на подготовленном нами наборе данных, что эта доля заметно растет в отличие от действительно стагнирующих средних значений Fsp^3 (см. табл. 40, определения дескрипторов приведены в разд. 2.1.2–2.1.3). Таким образом, вышеупомянутый вывод из статьи группы Н. Шнайдер о стагнации средних значений Fsp^3 из-за увеличения “жесткости” молекул “встает с головы на ноги”. На самом деле доля sp^3-гибридизированных атомов углерода, если они находятся в циклах, растёт, в то время как доля sp^3-гибридизированных атомов углерода, входящих в состав длинных алкильных цепочек — падает. Наблюдаемое явление происходит из-за всё более активного использования в синтезе новых потенциальных лекарственных веществ структурных блоков (*building blocks*), содержащих насыщенные карбо- и гетероциклы, в особенности спиро-циклы. Последнее наглядно демонстрируется заметным ростом средних значений дескриптора SPIRO.

Таблица 40. Средние значения молекулярных дескрипторов, рассчитанные для структур, раскрытых в патентных записях ведущих фармацевтических компаний за разные годы

Дескриптор	**1950–1983**	**1984–1990**	**1991–1997**	**1998–2004**	**2005–2011**	**2012–2018**	**Фактор роста**
MCE-18	43.2	50.1	56.3	57.1	64.9	75.9	**1.12**
MW	345.4	382.1	424.7	430.6	453.0	471.8	1.07
AR	0.86	0.9	0.95	0.98	0.99	0.99	1.03
NAR	0.72	0.71	0.75	0.72	0.73	0.83	1.03
CHIRAL	0.51	0.55	0.53	0.48	0.49	0.61	1.04
SPIRO	0.017	0.031	0.027	0.025	0.048	0.061	**1.36**
NCSPTR	0.22	0.23	0.21	0.22	0.24	0.26	1.04
Q^1	17.5	19.6	21.8	22.7	24.9	26.8	1.09
Fsp^3	38.3	39.4	35.7	33.0	33.1	34.5	0.98
Число раскрытых структур							
N	**765**	**2179**	**2741**	**6521**	**8932**	**3094**	**1.64**

Основной причиной того, что структурный ландшафт химического пространства малых потенциальных лекарственных молекул радикально изменился, мы считаем фундаментальное сдвиги в понимании белок-белковых взаимодействий (ББВ), как класса молекулярных мишеней. За последние полтора десятилетия исследования ББВ значительно продвинулись и привели к выявлению множества таргетируемых взаимодействий, включая p53/MDM2 и Bcl-2, GLP1R, Hsp70, NS5A, FAK, β-катенин и BRD. Многие авторы отмечают, что ББВ выходят далеко за рамки преобладающей парадигмы дизайна лекарств, поскольку молекулы, способные нарушать такие взаимодействия, существенно отличаются от классических хемотипов [192–194]. В последние годы было разработано большое количество инновационных синтетических каркасов с заметной трёхмерной сложностью, что позволило учёным создавать нетривиальные молекулярные структуры с высокой степенью пространственного разнообразия в качестве потенциальных ингибиторов ББВ.

С точки зрения классической медицинской химии новые соединения должны удовлетворять ряду тривиальных критериев, например “правилу пяти” Липински, чтобы достичь успеха в клинических испытаниях в качестве пероральных кандидатов. Однако ингибиторы ББВ зачастую значительно крупнее и массивнее традиционных лекарств [194]. С другой стороны, показательные примеры успешно выведенных на рынок или находящихся в разработке ингибиторов ББВ, которые не соответствуют параметрам RO5, продемонстрировали высокие значения MCE-18 (см рис. 65).

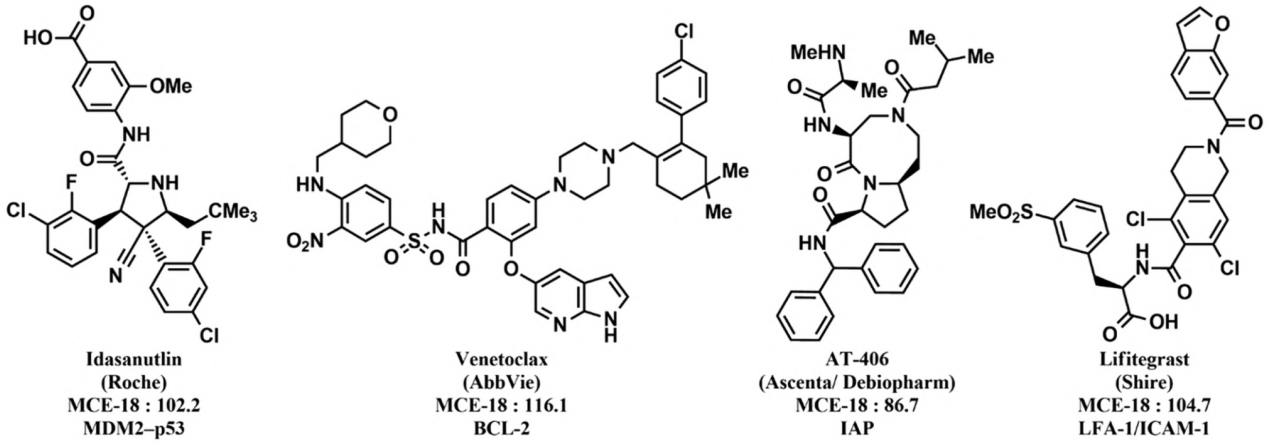


Рисунок 65. Показательные примеры ингибиторов ББВ, которые вышли на рынок или проходят клинические испытания.

В таких случаях медицинские химики сталкиваются с серьёзными вызовами и нуждаются в улучшении инструментов и методик для открытия новых лекарств за пределами правил Липински. Главная цель этих каркасов заключается в том, чтобы расположить присоединённые заместители так, чтобы достичь наилучшего взаимодействия с сайтом и соответствия «горячим точкам». Часто эти структурные блоки и скаффолды содержат сложную конфигурацию sp^3-атомов, формирующую пространственно предопределённое и жёсткое ядро с подходящими точками диверсификации. Учитывая, что простой индекс sp^3, соответствующий числу sp^3-гибридизованных атомов углерода, даёт высокий процент ложноположительных результатов, мы решили эту проблему с помощью дескриптора MCE-18 (см. 2.1.3).

Формула (1) отражает сложность всей молекулярной структуры, её качественные характеристики, в особенности, что касается содержания sp^3-гибридизированных атомов углерода. Как показано выше (табл. 39), квадратичный индекс возрастает с датой приоритета, присвоенной патентной записи, а также с датой вывода на рынок. Этот основанный на индексе Загреба показатель кодирует степень разветвлённости структурного каркаса через степень вершин и косвенно учитывает двумерную сложность. Компонента уравнения AR была включена в уравнение, поскольку было показано, что фенилаланин, триптофан и тирозин чаще чем обычно встречаются в интерфейсах многих ББВ [195]. Кроме того, этот терм уравнения позволяет отделять разветвленные алифатические и циклоалифатические вещества (например, жирные кислоты, структуры на основе стероидов, макролиды и полисахариды), широко представленные в “старой химии”, от пула новых молекулярных структур. В совокупности все компоненты формулы существенно способствуют отражению эволюции медицинской химии

и изменению качества sp^3-гибридизированных атомов углерода как с теоретической, так и со статистической точек зрения.

3.3.3. Дескриптор MCE-18 как альтернатива дескриптору Fsp^3 для анализа структурных трендов в медицинской химии

Поскольку дескриптор Fsp^3 ассоциируется в профессиональном сообществе с потенциальным успехом молекулярной структуры в клинических испытаниях [61] мы в решили сравнить способности дескрипторов MCE-18 и Fsp^3 моделировать структурные тренды на наборе данных о молекулярных структурах, собранных из патентов крупнейших фармацевтических компаний. Было показано (см. рис. 61), что дескриптор Fsp^3 (средневзвешенное за год значение для запатентованных в указанном году молекулярных структур) не чувствителен к очевидным структурным изменениям, которые можно наблюдать в особенности в последние годы, в то время как аналогично найденное средневзвешенное за год значение дескриптора MCE-18 моделирует восходящий тренд, который соответствует структурному усложнению представителей химического пространства, патентуемых фармацевтическими гигантами [196]. Показательно, что возрастание структурной сложности молекулярных структур потенциальных лекарственных веществ сопровождается падением научно-практической активности фармацевтических компаний, что говорит о фокусе на качество новых молекулярных структур, чем на их количество (см. гистограмму на фоне, рис. 66).

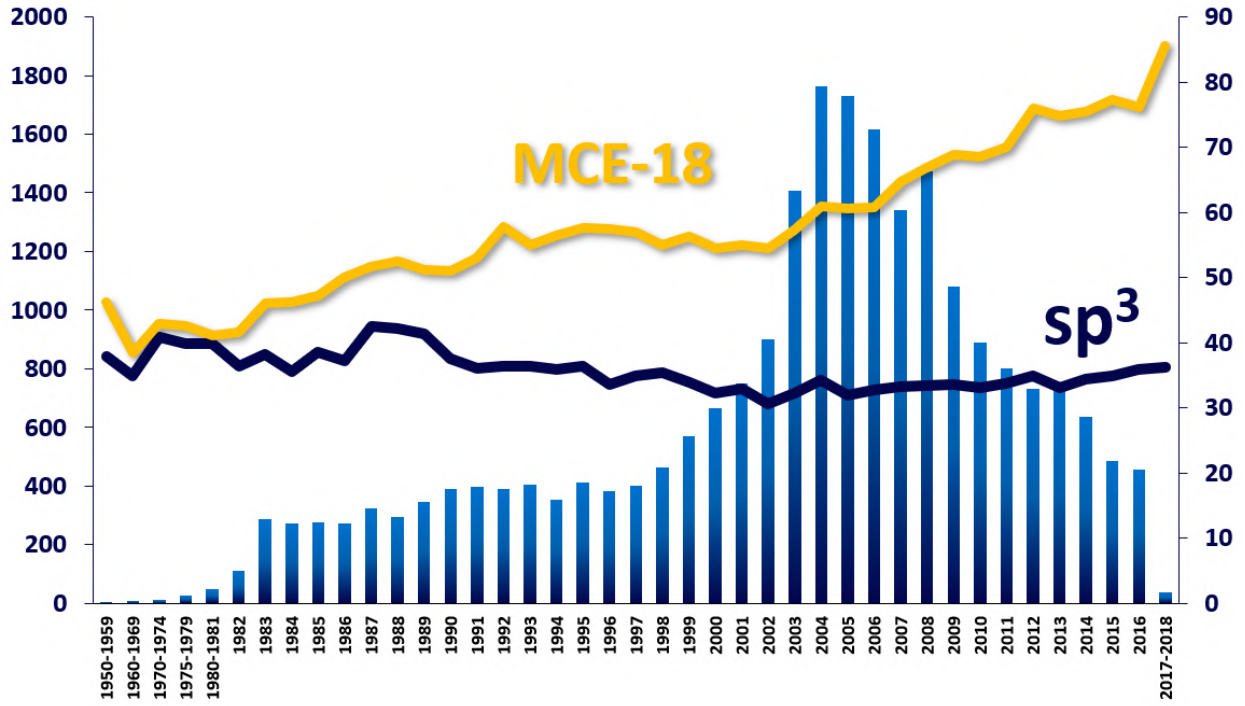


Рисунок 66. Сравнение структурных дескрипторов MCE-18 (желтая ломаная) и Fsp^3 (темно-синяя ломаная) на временной шкале (горизонтальная ось). Значения дескрипторов даны на правой вертикальной шкале, а на левой вертикальной шкале приведено количество патентов фармацевтических компаний, нормализованное относительно уникальных молекулярных структур потенциальных лекарственных веществ. Соответствующее нормализованное число патентов иллюстрирует научно-практическую деятельность 23 крупнейших фармацевтических компаний.

Последнее средневзвешенное за год значение дескриптора MCE-18 рассчитано для 2016 года и составляет 76.4. Тем не менее, реальный разброс значений остается высоким и в последние годы (статистика за 2017–2018 годы не была полностью учтена в связи с нераскрытыми данными о молекулярных структурах в патентах). Так, для всех основных семейств молекулярных мишеней лекарств было показано, что в последние 10 лет возросло количество

молекулярных структур со значением дескриптора MCE-18 более 100 единиц (см. рис. 67).

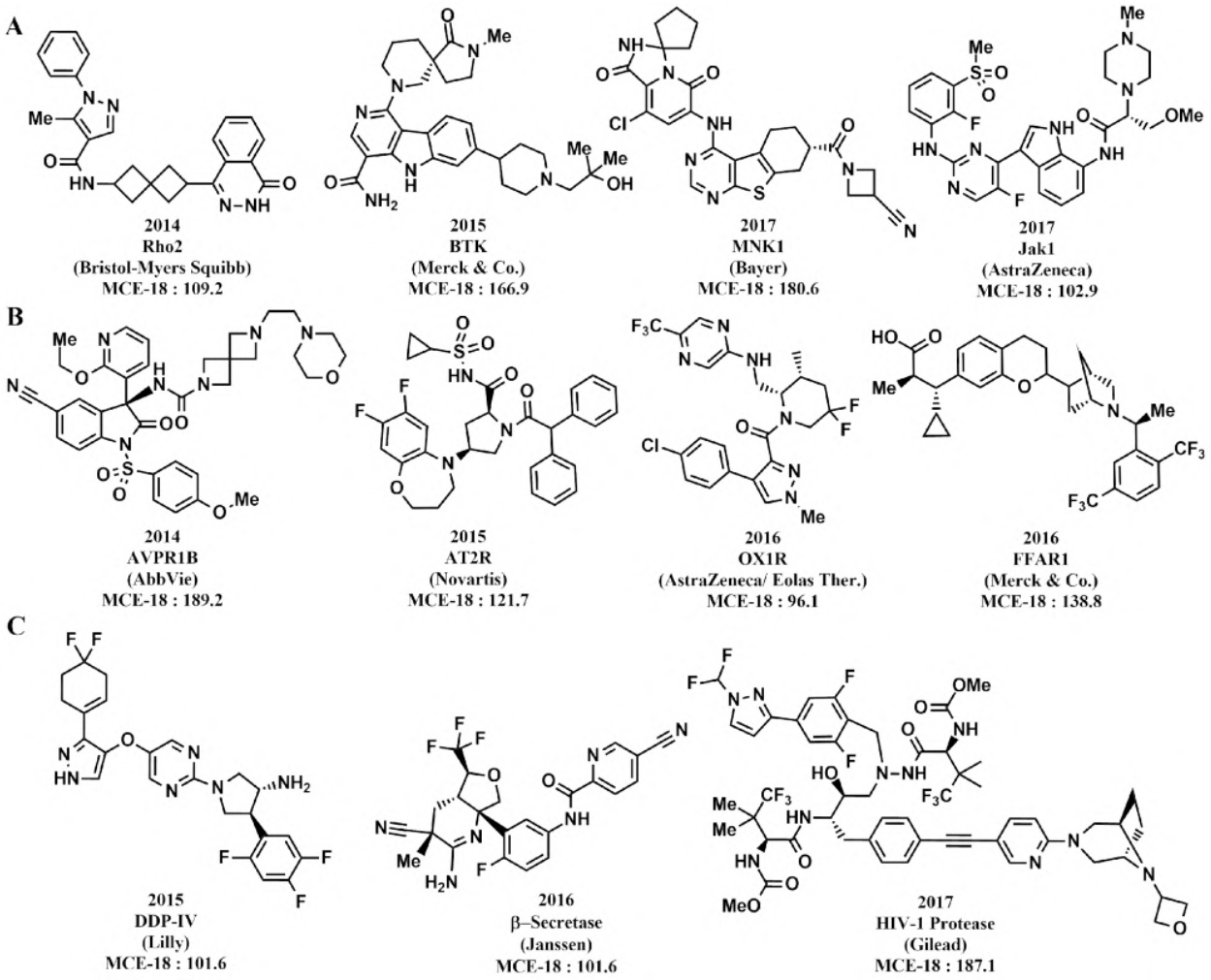


Рисунок 67. Типичные примеры структур с высокими значениями MCE-18, заявленные в патентах крупными фармацевтическими компаниями за последние 10 лет: **A)** ингибиторы протеиновых киназ, **B)** лиганды GPCR, **C)** ингибиторы протеаз.

Установка диапазона по умолчанию 75–150 для структурного дескриптора MCE-18 позволяет направлять генеративные алгоритмы платформы Chemistry42 создавать молекулярные структуры, соответствующие трендам, задаваемых крупнейшими фармацевтическими компаниями. Так, для демонстрации того, что тенденция дескриптора MCE-18 к увеличению его средних значений не зависит от конкретных фармацевтических компаний, графики зависимости средних значений дескриптора MCE-18 от времени были построены для 4 ведущих фармацевтических компаний (Sanofi, Pfizer, Merck&Co и AstraZeneca) по отдельности (см. рис. 68).

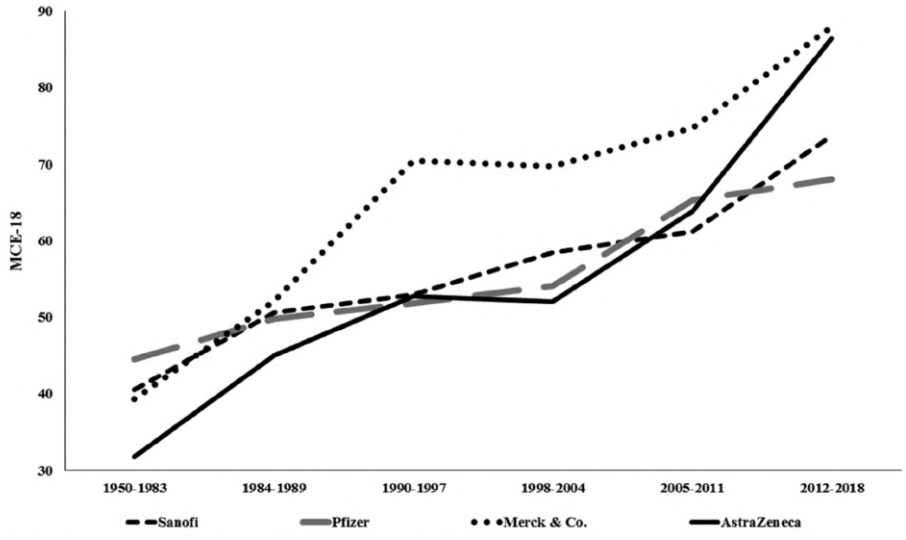


Рисунок 68. Графики зависимости средних значений MCE-18 от времени для препаратов, патентуемых 4 ведущими фармацевтическими компаниями.

Более того, было продемонстрировано, что тенденция средних значений дескриптора MCE-18 к росту также не зависит и от класса молекулярных мишеней лекарств (см. рис. 69). Средние значения MCE-18 растут как для киназных ингибиторов, так и для лигандов GPCR, несмотря на то что количество новых патентных записей для последних резко сократилось в последние годы. Особо же выраженный восходящий тренд наблюдается для протеазных ингибиторов, что связано с появлением обширного класса структурно сложных ингибиторов вирусных протеаз, преимущественно протеазы NS3/4A вируса гепатита С.

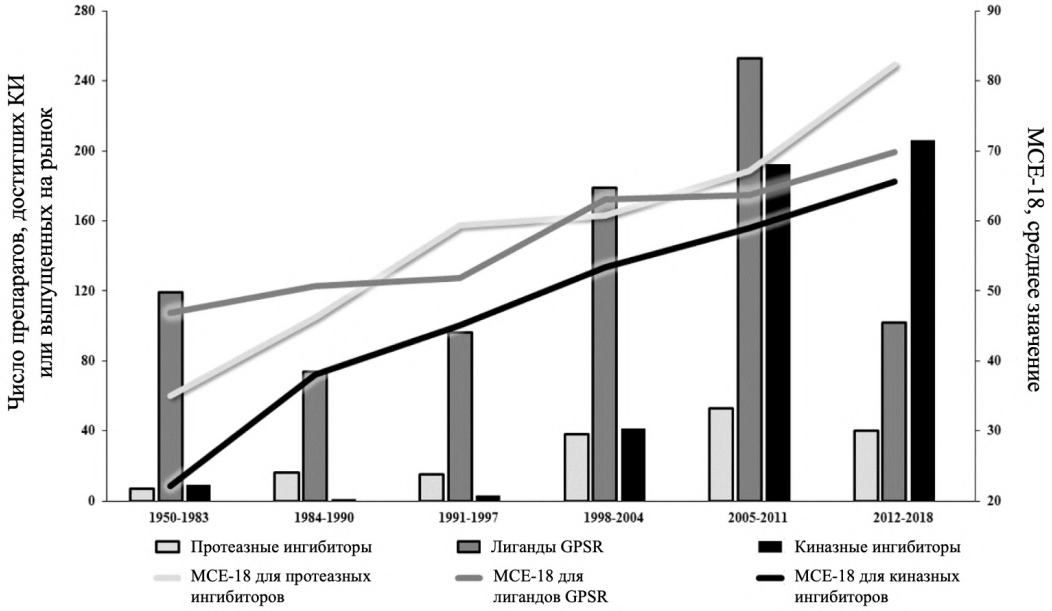


Рисунок 69. Количество препаратов (гистограмма), отсортированных по дате достижения контрольного этапа — клинических испытаний или вывода на рынок, в сравнении с непрерывным ростом средних значений MCE-18 (ломаные линии).

Также представляло интерес изучить вопрос, касающийся соотношения средних значений нового дескриптора с фазой разработки малых лекарственных молекул подобно тому, как был ранее введен в практику дескриптор Fsp^3 [61]. Поскольку подготовленный набор данных содержал такую информацию, то представлялась возможность провести подобное исследование (см. табл. 41).

Таблица 41. Средние значения молекулярных дескрипторов, рассчитанные для структур, раскрытых в патентных записях ведущих фармацевтических компаний и достигших этапов разработки лекарственных молекул (ДКИ — фаза доклинических испытаний, Фазы I–III — фазы клинических испытаний, Рынок — вышедшие на рынок зарегистрированные молекулы)

Дескриптор	**ДКИ**	**Фаза I**	**Фаза II**	**Фаза III**	**Рынок**	**Фактор роста**
MCE-18	57.3	56.2	55.9	54.0	45.98	0.95
MW	418.0	412.9	402.4	397.0	366.70	0.97
AR	0.91	0.87	0.87	0.85	0.80	0.97
NAR	0.72	0.72	0.71	0.71	0.67	0.98
CHIRAL	0.50	0.55	0.59	0.61	0.56	1.03
SPIRO	0.027	0.020	0.020	0.030	0.010	0.89
NCSPTR	0.22	0.24	0.24	0.24	0.23	1.01
Q^1	22.5	21.8	21.3	21.0	17.98	0.95
Fsp^3	36.4	40.5	41.5	41.8	44.65	1.05
Число раскрытых структур						
N	**38338**	**1678**	**1837**	**464**	**1370**	

В ходе анализа была подтверждена ранее обнаруженная тенденция малых молекул, прошедших клинические испытания, аккумулировать большую долю sp^3-гибридизованных атомов углерода. Что касается дескриптора MCE-18, то его средние значения не демонстрируют роста по мере продвижения к пространству вышедших на рынок молекул. Такая, на первый взгляд парадоксальная ситуация, объясняется как раз фундаментальным сдвигом в структурном ландшафте, его эволюцией, которая наблюдается в последнее десятилетие. Структурно сложные и крупные молекулярные структуры начали активно появляться как раз в последнее десятилетие, однако ещё не успели выйти на рынок, а те, что успели пройти одобрение регуляторами, пока не вносят ощутимого вклада в средние значения дескрипторов на фоне всех лекарств, большая часть которых представляет “старую химию”. Здесь особо стоит обратить внимание на резкую разницу в средних значениях дескриптора SPIRO между Фазой III (0.03) и Рынком (0.01) — сложные молекулярные структуры, включающие в себя спиро-сочлененные структурные блоки, содержатся в молекулах, находящихся в активной клинической разработке, и рынок ещё не приобрел их вклад. В этом смысле, нисходящий тренд дескриптора MCE-18 по мере разработки молекул, если учитывать

долгие сроки разработки, на самом деле указывает на то, что молекулы усложняются с каждым годом и молекулярные структуры, успевшие достигнуть Фазы I, сложнее тех, что успели достигнуть Фазы II, а последние сложнее тех, что успели достигнуть Фазы III. Таким образом, именно дескриптор MCE-18 в совокупности и его отдельные компоненты (MW, SPIRO, Q^1) позволяют продемонстрировать то, что невозможно показать при помощи дескриптора Fsp^3: успешность прохождения клинических испытаний малыми молекулами на длинной дистанции на самом деле не зависит каузально от Fsp^3, поскольку само химическое пространство зарегистрированных малых лекарственных молекул со структурной точки зрения радикально изменится за счет структурно сложным молекул, находящихся в клинической разработке, и структурно отличных от большинства зарегистрированных молекул. Продемонстрированная Ф. Ловерингом [61] зависимость представляет собой как раз пример ситуации, когда корреляция между некоторым дескриптором и трендом не означает наличие причинно-следственной связи. В противовес, мы видим, что анализ при помощи дескриптора MCE-18 и его компонент дает сущностное понимание эволюции химического пространства и её каузальности. Сам дескриптор MCE-18 представляет собой мощный инструмент для анализа тенденций, задаваемых ведущими фармацевтическими компаниями, которые можно наблюдать в изменении ландшафта химического пространства.

Заключение

В результате проведенного исследования по созданию модулей и модельных экспериментов для платформы генеративной химии, были выполнены все поставленные задачи и можно сделать следующие выводы:

1. На примере модельных экспериментов продемонстрирована способность разработанной платформы генеративной химии проводить *in silico* эксперименты согласно базовым сценариям молекулярного моделирования потенциальных лекарственных веществ.
2. Платформа генеративной химии была многократно валидирована экспериментально в рамках биологических тестирований и клинических испытаний. Польза платформы в разработке потенциальных лекарственных подтверждена успешным прохождением полноценных доклинических испытаний и IIa фазы клинических исследований для ингибитора TNIK киназы и I фазы клинических испытаний ингибитора главной протеазы коронавируса SARS-CoV-2.
3. Для платформы генеративной химии разработан метод моделирования синтетической доступности ReRSA, сочетающий лучшие практики оценки синтезируемости. *In silico* валидация алгоритма ReRSA выявила значимую корреляцию значений ReRSA с ретросинтетической сходимостью, получаемой в ходе работы автоматизированного ретросинтетического движка.
4. Разработан дескриптор MCE-18, эффективно моделирующий соответствие генерируемых платформой молекулярных структур структурным трендам медицинской химии на основе данных из патентов крупнейших фармацевтических компаний.

В связи с поступлением обратной связи от пользователей для следующей версии модуля ReRSA, выпуск которого намечен на 2026 год, были приоритезированы следующие функциональные изменения:

1. Предоставление пользователю возможности использовать пользовательский набор данных для переобучения статистической модели ReRSA;
2. Предоставление пользователю возможности использовать пользовательский набор КДИС;
3. Предоставление доступа к визуализации информации о статистическом анализе синтоноподобных фрагментов, покрывающих целевую молекулярную структуру, в

виде “тепловой карты”, где более встречаемые фрагменты будут окрашены в тёплые тона, а редко-встречаемые фрагменты — в холодные;

4. Расширение функционала модуля, ответственного за фильтрацию синтетически нерелевантных фрагментов, для покрытия шестичленных ароматических циклов, а также ароматических конденсированных систем 5–5, 5-6, 6-6; помимо этого планируется переработка и расширение азбуки элементов и заместителей для покрытия всех релевантных в медицинской химии случаев, а также визуализации — подсветки синтетически релевантной подструктуры в пользовательском интерфейсе.

Первые два упомянутых нововведения должны серьезно улучшить взаимодействие с пользователем для обеспечения его локальных потребностей исходя из его собственных ресурсов (например, набору КДИС из сервисной исследовательской организации, в которой пользователь заказывает проведение работ по синтезу) и/или исходя из нужд кампании по разработке определенного хемотипа, не представленного в исходном референсном наборе данных. Обновление под п.3 касается более “продвинутой” визуализации и должно предоставить пользователю больше понимания того, на чём основывается метод ReRSA.

В настоящий момент в ближайшей новой версии платформы не планируется радикальных обновлений базового алгоритма метода, тем не менее, планируется расширение функционала метода за счет покрытия большего числа фильтруемых, при помощи больших библиотек SMARTS-строк, типов гетероциклических систем (см. п. 4) для решения проблем, связанных со стереохимией, в настоящий момент времени мы работаем над созданием проприетарной библиотеки хемоинформатических инструментов MolKit, которая может поддерживать корректные стереохимические трансформации. При наличии технической возможности перевести часть функционала BRICSDecompose и RunReactants на основу пакета MolKit, стереохимическая информация может обрабатываться более корректным образом. Помимо этого, интерес представляет возможность выполнения ступенчатого ретросинтеза, когда продукты конвертации синтоноподобных фрагментов в синтетические эквиваленты в случае невозможности обнаружить последние в базе КДИС, будут сами подвергнуты квази-ретро синтетической фрагментации, подобной той, что используется в дефолтном алгоритме ReRSA. Конечные результаты такой ступенчатой квази-ретросинтетической процедуры могут быть затем конкатенированы в объединенный результат.

Что касается дескриптора MCE-18, то дальнейшее развитие дескриптора в том виде, в каком он был создан, с точки зрения добавления новых термов, не представляется возможным. В то же время моделирование трендов медицинской химии всегда представляется актуальной задачей и почти спустя 6 лет с начала работ по дескриптору MCE-18, как минимум обновление

обучающей выборки было бы полезным. Принципиально новое измерение будущий дескриптор мог бы приобрести, если бы имелась возможность анализировать структуры запатентованных молекулярных структур более тонкими инструментами, чем молекулярные дескрипторы, входящие в формулу MCE-18. Таким инструментом представляется полноподструктурный метод анализа, основы которого были заложены в рамках работы над настоящей диссертационной работой в приложении к анализу химического пространства 5-членных ароматических гетероциклов. В настоящее время нами ведется разработка методологии по анализу структурных трендов медицинской химии с использованием полноподструктурного метода анализа.

Благодарности

Автор диссертации выражает глубокую признательность научным руководителям Палюлину Владимиру Александровичу и Иваненкову Яну Андреевичу за чуткое руководство, гибкость и ценные советы в ходе работы над исследованиями. Автор выражает особую благодарность руководству группы компаний Insilico Medicine в лице президента компании Александра Мироновича Алипера и генерального директора Александра Александровича Жаворонкова за возможность использования коммерческих материалов в рамках платформы Chemistry42 для написания диссертации. Автор выражает особую признательность и благодарность Бондареву Никите Евгеньевичу (Insilico Medicine AI Limited) за ценную помощь в редактуре диссертации и совместное плодотворное творчество над методом моделирования синтетической доступности ReRSA. Автор выказывает глубокое уважение за профессионализм Федорченко Сергею Андреевичу (Insilico Medicine Hong Kong Ltd) в деле реализации программного кода метода моделирования синтетической доступности ReRSA. Автор благодарит за поддержку свою семью и друзей.

Список литературы

1. Mannhold R., Buschmann H., Holenz J. RNA as a drug target / под ред. Schneekloth J., Pettersson M. Weinheim, Germany: Wiley-VCH Verlag — 2024. — 416 с.

2. Zhang K., Yang X., Wang Y., Yu Y., Huang N., Li G., Li X., Wu J.C., Yang S. Artificial intelligence in drug development // Nature Medicine. — 2025. — Т. 31. — № 1. — С. 45–59.

3. Reymond J.-L. The chemical space project // Accounts of Chemical Research. — 2015. — Т. 48. — № 3. — С. 722–730.

4. van Tilborg D., Alenicheva A., Grisoni F. Exposing the Limitations of Molecular Machine Learning with Activity Cliffs // Journal of Chemical Information and Modeling. — 2022. — Т. 62. — № 23. — С. 5938–5951.

5. Stumpfe D., Hu H., Bajorath J. Evolving Concept of Activity Cliffs // ACS Omega. — 2019. — Т. 4. — № 11. — С. 14360–14368.

6. WHO Model List of Essential Medicines — 23rd list, 2023 [Электронный ресурс]. Режим доступа: https://www.who.int/publications/i/item/WHO-MHP-HPS-EML-2023.02.

7. Protein Data Bank [Электронный ресурс] Режим доступа: https://www.rcsb.org/.

8. Jumper J., Evans R., Pritzel A., Green T., Figurnov M., Ronneberger O., Tunyasuvunakool K., Bates R., Žídek A., Potapenko A., Bridgland A., Meyer C., Kohl S.A.A., Ballard A.J., Cowie A., Romera-Paredes B., Nikolov S., Jain R., Adler J., Back T., Petersen S., Reiman D., Clancy E., Zielinski M., Steinegger M., Pacholska M., Berghammer T., Bodenstein S., Silver D., Vinyals O., Senior A.W., Kavukcuoglu K., Kohli P., Hassabis D. Highly accurate protein structure prediction with AlphaFold // Nature. — 2021. — Т. 596. — № 7873. — С. 583–589.

9. Иваненков Я.А., Евтеев С.А., Малышев А.С., Терентьев В.А., Безруков Д.С., Ерещенко А.В., Корженевская А.А., Загрибельный Б.А., Шегай П.В., Каприн А.Д. AlphaFold в арсенале современного медицинского химика // Успехи химии. — Т. 93. — № 3. — С. RCR5107. (Переводная версия: Ivanenkov Y., Evteev S., Malyshev A.S., Terentiev V., Bezrukov D.S., Ereshchenko A.V., Korzhenevskaya A.A., Zagribelnyy B.A., Shegai P., Kaprin A. AlphaFold for a medicinal chemist: Tool or toy? // Russian Chemical Reviews. — 2024. — Т. 93. — № 3. — С. RCR5107.)

10. De novo Molecular Design / под ред. Schneider G. Weinheim, Germany: Wiley-VCH Verlag — 2014. — 576 с.

11. Bergner A., Parel S.P. Hit expansion approaches using multiple similarity methods and virtualized query structures // Journal of Chemical Information and Modeling. — 2013. — Т. 53. — № 5. — С. 1057–1066.

12. Fragment-based drug discovery: Lessons and outlook / под ред. Erlanson D.A., Jahnke W. Weinheim, Germany: Wiley-VCH Verlag — 2016. — 528 с.

13. Scaffold hopping in medicinal chemistry / под ред. Brown N., Mannhold R., Kubinyi H., Folkers G. Wiley-VCH Verlag — 2014. — 350 с.

14. Bajorath J. Comprehensive analysis of R-groups in medicinal chemistry // Future Medicinal Chemistry. — 2022. — Т. 14. — № 1. — С. 5–7.

15. Grenier D., Audebert S., Preto J., Guichou J.-F., Krimm I. Linkers in fragment-based drug design: an overview of the literature // Expert Opinion on Drug Discovery. — 2023. — Т. 18. — № 9. — С. 987–1009.

16. Bemis T.A., La Clair J.J., Burkart M.D. Unraveling the Role of Linker Design in Proteolysis Targeting Chimeras // Journal of Medicinal Chemistry. — 2021. — Т. 64. — № 12. — С. 8042–8052.

17. Brudy C., Walz C., Spiske M., Dreizler J.K., Hausch F. The Missing Link(er): A Roadmap to Macrocyclization in Drug Discovery // Journal of Medicinal Chemistry. — 2024. — Т. 67. — № 17. — С. 14768–14785.

18. Bian Y., Xie X.-Q. Generative chemistry: drug discovery with deep learning generative models // Journal of Molecular Modeling. — 2021. — Т. 27. — № 3. — С. 71.

19. Zhavoronkov A., Vanhaelen Q., Oprea T.I. Will Artificial Intelligence for Drug Discovery Impact Clinical Pharmacology? // Clinical Pharmacology and Therapeutics. — 2020. — Т. 107. — № 4. — С. 780–785.

20. Vanhaelen Q., Lin Y.-C., Zhavoronkov A. The Advent of Generative Chemistry // ACS Medicinal Chemistry Letters. — 2020. — Т. 11. — № 8. — С. 1496–1505.

21. Ivanenkov Y., Zagribelnyy B., Malyshev A., Evteev S., Terentiev V., Kamya P., Bezrukov D., Aliper A., Ren F., Zhavoronkov A. The Hitchhiker’s Guide to Deep Learning Driven Generative Chemistry // ACS Medicinal Chemistry Letters. — 2023. — Т. 14. — № 7. — С. 901–915.

22. Goldman B., Kearnes S., Kramer T., Riley P., Walters W.P. Defining Levels of Automated

Chemical Design // Journal of Medicinal Chemistry. — 2022. — Т. 65. — № 10. — С. 7073–7087.

23. Sheridan R.P., Kearsley S.K. Using a Genetic Algorithm To Suggest Combinatorial Libraries // Journal of chemical information and computer sciences. — 1995. — Т. 35. — № 2. — С. 310–320.

24. Rotstein S.H., Murcko M.A. GenStar: a method for de novo drug design // Journal of Computer-Aided Molecular Design. — 1993. — Т. 7. — № 1. — С. 23–43.

25. Lewell X.Q., Judd D.B., Watson S.P., Hann M.M. RECAP--retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry // Journal of Chemical Information and Computer Sciences. — 1998. — Т. 38. — № 3. — С. 511–522.

26. Degen J., Wegscheid-Gerlach C., Zaliani A., Rarey M. On the art of compiling and using «drug-like» chemical fragment spaces // ChemMedChem. — 2008. — Т. 3. — № 10. — С. 1503–1507.

27. Landrum G. RDKit: Open-source cheminformatics [Электронный ресурс]. Режим доступа: http://www.rdkit.org.

28. LeCun Y., Bengio Y., Hinton G. Deep learning // Nature. — 2015. — Т. 521. — № 7553. — С. 436–444.

29. Mamoshina P., Vieira A., Putin E., Zhavoronkov A. Applications of Deep Learning in Biomedicine // Molecular Pharmaceutics. — 2016. — Т. 13. — № 5. — С. 1445–1454.

30. Gasteiger J., Engel T. Chemoinformatics: A Textbook. John Wiley & Sons — 2006. — 680 с.

31. Gómez-Bombarelli R., Wei J.N., Duvenaud D., Hernández-Lobato J.M., Sánchez-Lengeling B., Sheberla D., Aguilera-Iparraguirre J., Hirzel T.D., Adams R.P., Aspuru-Guzik A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules // ACS Central Science. — 2018. — Т. 4. — № 2. — С. 268–276.

32. Polykovskiy D., Zhebrak A., Vetrov D., Ivanenkov Y., Aladinskiy V., Mamoshina P., Bozdaganyan M., Aliper A., Zhavoronkov A., Kadurin A. Entangled Conditional Adversarial Autoencoder for de Novo Drug Discovery // Molecular Pharmaceutics. — 2018. — Т. 15. — № 10. — С. 4398–4405.

33. Putin E., Asadulaev A., Vanhaelen Q., Ivanenkov Y., Aladinskaya A.V., Aliper A., Zhavoronkov A. Adversarial Threshold Neural Computer for Molecular de Novo Design // Molecular Pharmaceutics. — 2018. — Т. 15. — № 10. — С. 4386–4397.

34. Putin E., Asadulaev A., Ivanenkov Y., Aladinskiy V., Sanchez-Lengeling B., Aspuru-Guzik A., Zhavoronkov A. Reinforced Adversarial Neural Computer for de Novo Molecular Design // Journal of Chemical Information and Modeling. — 2018. — Т. 58. — № 6. — С. 1194–1204.

35. Zhavoronkov A., Ivanenkov Y.A., Aliper A., Veselov M.S., Aladinskiy V.A., Aladinskaya A.V., Terentiev V.A., Polykovskiy D.A., Kuznetsov M.D., Asadulaev A., Volkov Y., Zholus A., Shayakhmetov R.R., Zhebrak A., Minaeva L.I., Zagribelnyy B.A., Lee L.H., Soll R., Madge D., Xing L., Guo T., Aspuru-Guzik A. Deep learning enables rapid identification of potent DDR1 kinase inhibitors // Nature Biotechnology. — 2019. — Т. 37. — № 9. — С. 1038–1040.

36. Kohonen T. Self-Organizing Maps. Third, Extended Edition. Berlin, Germany: Springer-Verlag — 2001.

37. Chemistry42 [Электронный ресурс]. Режим доступа: https://insilico.com/chemistry42.

38. Makya [Электронный ресурс]. Режим доступа: https://iktos.ai/solution/makya.

39. Spaya [Электронный ресурс]. Режим доступа: https://iktos.ai/solution/spaya.

40. Bos P.H., Houang E.M., Ranalli F., Leffler A.E., Boyles N.A., Eyrich V.A., Luria Y., Katz D., Tang H., Abel R., Bhat S. AutoDesigner, a De Novo Design Algorithm for Rapidly Exploring Large Chemical Space for Lead Optimization: Application to the Design and Synthesis of d-Amino Acid Oxidase Inhibitors // Journal of Chemical Information and Modeling. — 2022. — Т. 62. — № 8. — С. 1905–1915.

41. Tang H., Jensen K., Houang E., McRobb F.M., Bhat S., Svensson M., Bochevarov A., Day T., Dahlgren M.K., Bell J.A., Frye L., Skene R.J., Lewis J.H., Osborne J.D., Tierney J.P., Gordon J.A., Palomero M.A., Gallati C., Chapman R.S.L., Jones D.R., Hirst K.L., Sephton M., Chauhan A., Sharpe A., Tardia P., Dechaux E.A., Taylor A., Waddell R.D., Valentine A., Janssens H.B., Aziz O., Bloomfield D.E., Ladha S., Fraser I.J., Ellard J.M. Discovery of a Novel Class of d-Amino Acid Oxidase Inhibitors Using the Schrödinger Computational Platform // Journal of Medicinal Chemistry. — 2022. — Т. 65. — № 9. — С. 6775–6802.

42. Jones J., Clark R.D., Lawless M.S., Miller D.W., Waldman M. The AI-driven Drug Design (AIDD) platform: an interactive multi-parameter optimization system integrating molecular

evolution with physiologically based pharmacokinetic simulations // Journal of Computer-Aided Molecular Design. — 2024. — Т. 38. — № 1. — С. 14.

43. Ertl P., Schuffenhauer A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions // Journal of Cheminformatics. — 2009. — Т. 1. — № 1. — С. 8.

44. Rusinko A., Rezaei M., Friedrich L., Buchstaller H.-P., Kuhn D., Ghogare A. AIDDISON: Empowering Drug Discovery with AI/ML and CADD Tools in a Secure, Web-Based SaaS Platform // Journal of Chemical Information and Modeling. — 2024. — Т. 64. — № 1. — С. 3–8.

45. Brocklehurst C.E., Altmann E., Bon C., Davis H., Dunstan D., Ertl P., Ginsburg-Moraff C., Grob J., Gosling D.J., Lapointe G., Marziale A.N., Mues H., Palmieri M., Racine S., Robinson R.I., Springer C., Tan K., Ulmer W., Wyler R. MicroCycle: An Integrated and Automated Platform to Accelerate Drug Discovery // Journal of Medicinal Chemistry. — 2024. — Т. 67. — № 3. — С. 2118–2128.

46. Ghiandoni G.M., Evertsson E., Riley D.J., Tyrchan C., Rathi P.C. Augmenting DMTA using predictive AI modelling at AstraZeneca // Drug Discovery Today. — 2024. — Т. 29. — № 4. — С. 103945.

47. Arnold C. Inside the nascent industry of AI-designed drugs // Nature Medicine. — 2023. — Т. 29. — № 6. — С. 1292–1295.

48. Jayatunga M.K.P., Xie W., Ruder L., Schulze U., Meier C. AI in small-molecule drug discovery: a coming wave? // Nature Reviews. Drug Discovery. — 2022. — Т. 21. — № 3. — С. 175–176.

49. Ren F., Aliper A., Chen J., Zhao H., Rao S., Kuppe C., Ozerov I.V., Zhang M., Witte K., Kruse C., Aladinskiy V., Ivanenkov Y., Polykovskiy D., Fu Y., Babin E., Qiao J., Liang X., Mou Z., Wang H., Pun F.W., Ayuso P.T., Veviorskiy A., Song D., Liu S., Zhang B., Naumov V., Ding X., Kukharenko A., Izumchenko E., Zhavoronkov A. A small-molecule TNIK inhibitor targets fibrosis in preclinical and clinical models // Nature Biotechnology. — 2024. — Т. 43. — С. 63–75.

50. Xu Z., Ren F., Wang P., Cao J., Tan C., Ma D., Zhao L., Dai J., Ding Y., Fang H., Li H., Liu H., Luo F., Meng Y., Pan P., Xiang P., Xiao Z., Rao S., Satler C., Liu S., Lv Y., Zhao H., Chen S., Cui H., Korzinkin M., Gennert D., Zhavoronkov A. A generative AI-discovered TNIK inhibitor

for idiopathic pulmonary fibrosis: a randomized phase 2a trial // Nature Medicine. — 2025. — T. 31. — № 8. — C. 2602–2610.

51. Aladinskiy V., Kruse C., Qin L., Babin E., Fan Y., Andreev G., Zhao H., Fu Y., Zhang M., Ivanenkov Y., Aliper A., Zhavoronkov A., Ren F. Discovery of bis-imidazolecarboxamide derivatives as novel, potent, and selective TNIK inhibitors for the treatment of idiopathic pulmonary fibrosis // Journal of Medicinal Chemistry. — 2024. — T. 67. — № 21. — C. 19121–19142.

52. Ertl P. Cheminformatics analysis of organic substituents: identification of the most common substituents, calculation of substituent properties, and automatic identification of drug-like bioisosteric groups // Journal of chemical information and computer sciences. — 2003. — T. 43. — № 2. — C. 374–380.

53. Dömling A. Protein-Protein Interactions in Drug Discovery. John Wiley & Sons — 2013. — 334 c.

54. Lipinski C.A., Lombardo F., Dominy B.W., Feeney P.J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings // Advanced Drug Delivery Reviews. — 2001. — T. 46. — № 1–3. — C. 3–26.

55. Doak B.C., Over B., Giordanetto F., Kihlberg J. Oral druggable space beyond the rule of 5: insights from drugs and clinical candidates // Chemistry & Biology. — 2014. — T. 21. — № 9. — C. 1115–1142.

56. Shultz M.D. Two decades under the influence of the rule of five and the changing properties of approved oral drugs // Journal of Medicinal Chemistry. — 2019. — T. 62. — № 4. — C. 1701–1714.

57. Schuffenhauer A., Schneider N., Hintermann S., Auld D., Blank J., Cotesta S., Engeloch C., Fechner N., Gaul C., Giovannoni J., Jansen J., Joslin J., Krastel P., Lounkine E., Manchester J., Monovich L.G., Pelliccioli A.P., Schwarze M., Shultz M.D., Stiefl N., Baeschlin D.K. Evolution of Novartis' small molecule screening deck design // Journal of Medicinal Chemistry. — 2020. — T. 63. — № 23. — C. 14425–14447.

58. O' Donovan D.H., De Fusco C., Kuhnke L., Reichel A. Trends in molecular properties, bioavailability, and permeability across the Bayer compound collection // Journal of Medicinal Chemistry. — 2023. — T. 66. — № 4. — C. 2347–2360.

59. DeGoey D.A., Chen H.-J., Cox P.B., Wendt M.D. Beyond the rule of 5: Lessons learned from AbbVie's drugs and compound collection // Journal of Medicinal Chemistry. — 2018. — Т. 61. — № 7. — С. 2636–2651.

60. Price E., Weinheimer M., Rivkin A., Jenkins G., Nijsen M., Cox P.B., DeGoey D. Beyond rule of five and PROTACs in modern drug discovery: Polarity reducers, chameleonicity, and the evolving physicochemical landscape // Journal of Medicinal Chemistry. — 2024. — Т. 67. — № 7. — С. 5683–5698.

61. Lovering F., Bikker J., Humblet C. Escape from flatland: increasing saturation as an approach to improving clinical success // Journal of Medicinal Chemistry. — 2009. — Т. 52. — № 21. — С. 6752–6756.

62. Baber J.C., Feher M. Predicting synthetic accessibility: application in drug discovery and development // Mini Reviews in Medicinal Chemistry. — 2004. — Т. 4. — № 6. — С. 681–692.

63. Coley C.W., Rogers L., Green W.H., Jensen K.F. SCScore: Synthetic Complexity Learned from a Reaction Corpus // Journal of Chemical Information and Modeling. — 2018. — Т. 58. — № 2. — С. 252–261.

64. Bertz S.H. The first general index of molecular complexity // Journal of the American Chemical Society. — 1981. — Т. 103. — № 12. — С. 3599–3601.

65. Barone R., Chanon M. A new and simple approach to chemical complexity. Application to the synthesis of natural products // Journal of Chemical Information and Computer Sciences. — 2001. — Т. 41. — № 2. — С. 269–272.

66. Todeschini R., Consonni V. Handbook of Molecular Descriptors. John Wiley & Sons — 2008. — 688 с.

67. Whitlock H.W. On the Structure of Total Synthesis of Complex Natural Products // The Journal of organic chemistry. — 1998. — Т. 63. — № 22. — С. 7982–7989.

68. Rücker G., Rücker C. Walk counts, labyrinthicity, and complexity of acyclic and cyclic graphs and molecules // Journal of chemical information and. — 2000. — Т. 40. — № 1. — С. 99–106.

69. Kochev N., Avramova S., Angelov P., Jeliazkova N. Computational prediction of synthetic accessibility of organic molecules with Ambit-synthetic accessibility tool // Organic Chemistry:

An Indian Journal — 2018. — Т. 14. — С. 123.

70. ASKCOS: Software tools for organic synthesis [Электронный ресурс]. Режим доступа: https://askcos.mit.edu/.

71. AiZynthFinder: A tool for retrosynthetic planning [Электронный ресурс]. Режим доступа: https://github.com/MolecularAI/aizynthfinder.

72. SYNTHIA Retrosynthesis Software [Электронный ресурс]. Режим доступа: https://www.sigmaaldrich.com/AE/en/services/software-and-digital-platforms/synthia-retrosynthesis-software.

73. CAS SciFinder - retrosynthesis software [Электронный ресурс]. Режим доступа: https://www.cas.org/solutions/cas-scifinder-discovery-platform/cas-scifinder/synthesis-planning.

74. Reaxys [Электронный ресурс]. Режим доступа: https://www.elsevier.com/solutions/reaxys.

75. Патент US20220172802A1. Retrosynthesis systems and methods: опубл. 02.06.2022. / Konstantinov A., Putin E.O., Zagribelnyy B., Ivanenkov Y.A., Zhavoronkovs A.

76. Takaoka Y., Endo Y., Yamanobe S., Kakinuma H., Okubo T., Shimazaki Y., Ota T., Sumiya S., Yoshikawa K. Development of a method for evaluating drug-likeness and ease of synthesis using a data set in which compounds are assigned scores based on chemists' intuition // Journal of Chemical Information and Computer Sciences. — 2003. — Т. 43. — № 4. — С. 1269–1275.

77. Neeser R.M., Correia B., Schwaller P. FSscore: A personalized machine learning-based synthetic feasibility score // Chemistry-Methods. — 2024. — Т. 4. — № 11. — С. e202400024

78. Voršilák M., Kolář M., Čmelo I., Svozil D. SYBA: Bayesian estimation of synthetic accessibility of organic compounds // Journal of Cheminformatics. — 2020. — Т. 12. — № 35.

79. Li B., Chen H. Prediction of Compound Synthesis Accessibility Based on Reaction Knowledge Graph // Molecules. — 2022. — Т. 27. — № 3. — С. 1039.

80. Thakkar A., Chadimová V., Bjerrum E.J., Engkvist O., Reymond J.-L. Retrosynthetic accessibility score (RAscore) - rapid machine learned synthesizability classification from AI driven retrosynthetic planning // Chemical Science . — 2021. — Т. 12. — № 9. — С. 3339–3349.

81. Podolyan Y., Walters M.A., Karypis G. Assessing synthetic accessibility of chemical compounds using machine learning methods // Journal of Chemical Information and Modeling. — 2010. — Т. 50. — № 6. — С. 979–991.

82. Yu J., Wang J., Zhao H., Gao J., Kang Y., Cao D., Wang Z., Hou T. Organic Compound Synthetic Accessibility Prediction Based on the Graph Attention Mechanism // Journal of Chemical Information and Modeling. — 2022. — Т. 62. — № 12. — С. 2973–2986.

83. Liu C.-H., Korablyov M., Jastrzębski S., Włodarczyk-Pruszyński P., Bengio Y., Segler M. RetroGNN: Fast Estimation of Synthesizability for Virtual Screening and De Novo Design by Learning from Slow Retrosynthesis Software // Journal of Chemical Information and Modeling. — 2022. — Т. 62. — № 10. — С. 2293–2300.

84. Wang S., Wang L., Li F., Bai F. DeepSA: a deep-learning driven predictor of compound synthesis accessibility // Journal of Cheminformatics. — 2023. — Т. 15. — № 103.

85. Kim H., Lee K., Kim C., Lim J., Kim W.Y. DFRscore: Deep Learning-Based Scoring of Synthetic Complexity with Drug-Focused Retrosynthetic Analysis for High-Throughput Virtual Screening // Journal of Chemical Information and Modeling. — 2024. — Т. 64. — № 7. — С. 2432–2444.

86. Huang Q., Li L.-L., Yang S.-Y. RASA: a rapid retrosynthesis-based scoring method for the assessment of synthetic accessibility of drug-like molecules // Journal of Chemical Information and Modeling. — 2011. — Т. 51. — № 10. — С. 2768–2777.

87. Parrot M., Tajmouati H., da Silva V.B.R., Atwood B.R., Fourcade R., Gaston-Mathé Y., Do Huu N., Perron Q. Integrating synthetic accessibility with AI-based generative drug design // Journal of Cheminformatics. — 2023. — Т. 15. — № 83.

88. Chen S., Jung Y. Estimating the synthetic accessibility of molecules with building block and reaction-aware SAScore // Journal of Cheminformatics. — 2024. — Т. 16. — № 83.

89. Патент US20230154572A1. Retrosynthesis-related synthetic accessibility: опубл. 18.05.2023. / Zagribelnyy B., Putin E.O., Fedorchenko S.A., Ivanenkov Y.A., Zavoronkovs A.

90. Corey E.J. General methods for the construction of complex molecules // Pure and Applied Chemistry. — 1967. — Т. 14. — № 1. — С. 19–38.

91. Johnson A.P., Marshall C., Judson P.N. Starting material oriented retrosynthetic analysis in the

LHASA program. 1. General description // Journal of chemical information and computer sciences. — 1992. — Т. 32. — № 5. — С. 411–417.

92. Gillet V.J., Myatt G., Zsoldos Z., Johnson A.P. SPROUT, HIPPO and CAESA: Tools for de novo structure generation and estimation of synthetic accessibility // Perspectives in Drug Discovery and Design. — 1995. — Т. 3. — № 1. — С. 34–50.

93. Saigiridharan L., Hassen A.K., Lai H., Torren-Peraire P., Engkvist O., Genheden S. AiZynthFinder 4.0: developments based on learnings from 3 years of industrial application // Journal of Cheminformatics. — 2024. — Т. 16. — № 57.

94. Mikulak-Klucznik B., Gołębiowska P., Bayly A.A., Popik O., Klucznik T., Szymkuć S., Gajewska E.P., Dittwald P., Staszewska-Krajewska O., Beker W., Badowski T., Scheidt K.A., Molga K., Mlynarski J., Mrksich M., Grzybowski B.A. Computational planning of the synthesis of complex natural products // Nature. — 2020. — Т. 588. — № 7836. — С. 83–88.

95. Allu T.K., Oprea T.I. Rapid evaluation of synthetic and molecular complexity for in silico chemistry // Journal of Chemical Information and Modeling. — 2005. — Т. 45. — № 5. — С. 1237–1243.

96. Coley C.W. Defining and Exploring Chemical Spaces // Trends in Chemistry. — 2021. — Т. 3. — № 2. — С. 133–145.

97. Mercado R., Kearnes S.M., Coley C.W. Data Sharing in Chemistry: Lessons Learned and a Case for Mandating Structured Reaction Data // Journal of Chemical Information and Modeling. — 2023. — Т. 63. — № 14. — С. 4253–4265.

98. Wigh D.S., Goodman J.M., Lapkin A.A. A review of molecular representation in the age of machine learning // Wiley Interdisciplinary Reviews. Computational Molecular Science. — 2022. — Т. 12. — № 5. — С. e1603

99. Sheridan R.P., Zorn N., Sherer E.C., Campeau L.-C., Chang C. (zhenyu), Cumming J., Maddess M.L., Nantermet P.G., Sinz C.J., O'Shea P.D. Modeling a Crowdsourced Definition of Molecular Complexity // Journal of chemical information and modeling. — 2014. — Т. 54. — № 6. — С. 1604–1616.

100. Fukunishi Y., Kurosawa T., Mikami Y., Nakamura H. Prediction of synthetic accessibility based on commercially available compound databases // Journal of Chemical Information and Modeling. — 2014. — Т. 54. — № 12. — С. 3259–3267.

101. Morgan H.L. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service // Journal of chemical documentation. — 1965. — Т. 5. — № 2. — С. 107–113.

102. 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.

103. Neeser R.M., Correia B., Schwaller P. FSscore: A Personalized Machine Learning-Based Synthetic Feasibility Score // Chemistry–Methods. — 2024. — Т.4. — С. e202400024.

104. Chapelle O., Schölkopf B., Zien A. Semi-supervised Learning. MIT Press — 2006. — 508 с.

105. Fourches D., Ash J. 4D- quantitative structure–activity relationship modeling: making a comeback // Expert opinion on drug discovery. — 2019. — Т. 14. — № 12. — С. 1227–1235.

106. Karniadakis G.E., Kevrekidis I.G., Lu L., Perdikaris P., Wang S., Yang L. Physics-informed machine learning // Nature Reviews Physics. — 2021. — Т. 3. — № 6. — С. 422–440.

107. Lajiness M.S., Maggiora G.M., Shanmugasundaram V. Assessment of the consistency of medicinal chemists in reviewing sets of compounds // Journal of Medicinal Chemistry. — 2004. — Т. 47. — № 20. — С. 4891–4896.

108. PubChem [Электронный ресурс]. Режим доступа: https://pubchem.ncbi.nlm.nih.gov/.

109. Flick A.C., Ding H.X., Leverett C.A., Kyne R.E. Jr, Liu K.K.-C., Fink S.J., O'Donnell C.J. Synthetic Approaches to the New Drugs Approved During 2015 // Journal of Medicinal Chemistry. — 2017. — Т. 60. — № 15. — С. 6480–6515.

110. Genheden S., Thakkar A., Chadimová V., Reymond J.-L., Engkvist O., Bjerrum E. AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning // Journal of Cheminformatics. — 2020. — Т. 12. — № 70.

111. Cortellis Drug Discovery Intelligence [Электронный ресурс] // Clarivate. — 2022. URL: https://clarivate.com/products/biopharma/research-development/pre-clinical-intelligence-analytics/ (дата обращения: 29.08.2024).

112. ChemDiv [Электронный ресурс]. Режим доступа: https://www.chemdiv.com/.

113. Balakin K.V. Pharmaceutical Data Mining: Approaches and Applications for Drug Discovery.

John Wiley & Sons — 2009. — 584 c.

114. Schneider N., Lowe D.M., Sayle R.A., Tarselli M.A., Landrum G.A. Big Data from Pharmaceutical Patents: A Computational Analysis of Medicinal Chemists' Bread and Butter // Journal of Medicinal Chemistry. — 2016. — Т. 59. — № 9. — С. 4385–4402.

115. Brown D.G., Boström J. Analysis of Past and Present Synthetic Methodologies on Medicinal Chemistry: Where Have All the New Reactions Gone? // Journal of Medicinal Chemistry. — 2016. — Т. 59. — № 10. — С. 4443–4458.

116. ChEMBL Database [Электронный ресурс]. Режим доступа: https://www.ebi.ac.uk/chembl/ (дата обращения: 17.08.2024).

117. Screening collections - Enamine [Электронный ресурс]. Режим доступа: https://enamine.net/compound-collections/screening-collection.

118. Enamine Building Blocks Catalog [Электронный ресурс]. Режим доступа: https://enamine.net/building-blocks/building-blocks-catalog.

119. Chemical Abstracts Service (CAS). SciFinder // Journal of the Medical Library Association. — 2018. — Т. 106. — № 4. — С. 558–590.

120. SMARTS^{T} — A Language for Describing Molecular Patterns [Электронный ресурс]. Режим доступа: https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html.

121. Balaban A.T. Chemical graphs // Theoretica Chimica Acta. — 1979. — Т. 53. — № 4. — С. 355–375.

122. Dömling A., Wang W., Wang K. Chemistry and biology of multicomponent reactions // Chemical Reviews. — 2012. — Т. 112. — № 6. — С. 3083–3135.

123. Marsault E., Peterson M.L. Macrocycles are great cycles: applications, opportunities, and challenges of synthetic macrocycles in drug discovery // Journal of Medicinal Chemistry. — 2011. — Т. 54. — № 7. — С. 1961–2004.

124. Walters W.P., Murcko M. Assessing the impact of generative AI on medicinal chemistry // Nature Biotechnology. — 2020. — Т. 38. — № 2. — С. 143–145.

125. Zhavoronkov A., Aspuru-Guzik A. Reply to «Assessing the impact of generative AI on medicinal chemistry» // Nature biotechnology. — 2020. — Т. 38. — № 2. — С. 146.

126. Ivanenkov Y.A., Polykovskiy D., Bezrukov D., Zagribelnyy B., Aladinskiy V., Kamya P., Aliper A., Ren F., Zhavoronkov A. Chemistry42: An AI-Driven Platform for Molecular Design and Optimization // Journal of Chemical Information and Modeling. — 2023. — Т. 63. — № 3. — С. 695–701.

127. Baell J.B., Holloway G.A. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays // Journal of Medicinal Chemistry. — 2010. — Т. 53. — № 7. — С. 2719–2740.

128. Liu Z., Su M., Han L., Liu J., Yang Q., Li Y., Wang R. Forging the basis for developing protein-ligand interaction Scoring Functions // Accounts of Chemical Research. — 2017. — Т. 50. — № 2. — С. 302–309.

129. Journal of Medicinal Chemistry [Электронный ресурс]. Режим доступа: https://pubs.acs.org/journal/jmcmar.

130. Osipiuk J., Azizi S.-A., Dvorkin S., Endres M., Jedrzejczak R., Jones K.A., Kang S., Kathayat R.S., Kim Y., Lisnyak V.G., Maki S.L., Nicolaescu V., Taylor C.A., Tesar C., Zhang Y.-A., Zhou Z., Randall G., Michalska K., Snyder S.A., Dickinson B.C., Joachimiak A. Structure of papain-like protease from SARS-CoV-2 and its complexes with non-covalent inhibitors // Nature Communications. — 2021. — Т. 12. — № 1. — С. 743.

131. PDB: 7CMD [Электронный ресурс]. Режим доступа: https://www.rcsb.org/structure/7CMD.

132. Assay and Activity Questions [Электронный ресурс]. Режим доступа: https://chembl.gitbook.io/chembl-interface-documentation/frequently-asked-questions/chembl-data-questions#what-is-the-confidence-score.

133. Chrencik J.E., Patny A., Leung I.K., Korniski B., Emmons T.L., Hall T., Weinberg R.A., Gormley J.A., Williams J.M., Day J.E., Hirsch J.L., Kiefer J.R., Leone J.W., Fischer H.D., Sommers C.D., Huang H.-C., Jacobsen E.J., Tenbrink R.E., Tomasselli A.G., Benson T.E. Structural and thermodynamic characterization of the TYK2 and JAK3 kinase domains in complex with CP-690550 and CMP-6 // Journal of Molecular Biology. — 2010. — Т. 400. — № 3. — С. 413–433.

134. PDB: 3LXK [Электронный ресурс]. Режим доступа: https://www.rcsb.org/structure/3lxk.

135. Kategaya L., Di Lello P., Rougé L., Pastor R., Clark K.R., Drummond J., Kleinheinz T., Lin E., Upton J.-P., Prakash S., Heideker J., McCleland M., Ritorto M.S., Alessi D.R., Trost M.,

Bainbridge T.W., Kwok M.C.M., Ma T.P., Stiffler Z., Brasher B., Tang Y., Jaishankar P., Hearn B.R., Renslo A.R., Arkin M.R., Cohen F., Yu K., Peale F., Gnad F., Chang M.T., Klijn C., Blackwood E., Martin S.E., Forrest W.F., Ernst J.A., Ndubaku C., Wang X., Beresini M.H., Tsui V., Schwerdtfeger C., Blake R.A., Murray J., Maurer T., Wertz I.E. USP7 small-molecule inhibitors interfere with ubiquitin binding // Nature. — 2017. — Т. 550. — № 7677. — С. 534–538.

136. PDB: 5UQX [Электронный ресурс]. Режим доступа: https://www.rcsb.org/structure/5UQX.

137. Eduful B.J., O'Byrne S.N., Temme L., Asquith C.R.M., Liang Y., Picado A., Pilotte J.R., Hossain M.A., Wells C.I., Zuercher W.J., Catta-Preta C.M.C., Zonzini Ramos P., Santiago A. de S., Couñago R.M., Langendorf C.G., Nay K., Oakhill J.S., Pulliam T.L., Lin C., Awad D., Willson T.M., Frigo D.E., Scott J.W., Drewry D.H. Hinge Binder Scaffold Hopping Identifies Potent Calcium/Calmodulin-Dependent Protein Kinase Kinase 2 (CAMKK2) Inhibitor Chemotypes // Journal of Medicinal Chemistry. — 2021. — Т. 64. — № 15. — С. 10849–10877.

138. PDB: 6BKU [Электронный ресурс]. Режим доступа: https://www.rcsb.org/structure/6BKU.

139. Lee Y., Kim H., Kim H., Cho H.Y., Jee J.-G., Seo K.-A., Son J.B., Ko E., Choi H.G., Kim N.D., Kim I. X-ray crystal structure-guided design and optimization of 7H-pyrrolo[2,3-d]pyrimidine-5-carbonitrile scaffold as a potent and orally active monopolar spindle 1 inhibitor // Journal of Medicinal Chemistry. — 2021. — Т. 64. — № 10. — С. 6985–6995.

140. PDB: 5N7V [Электронный ресурс]. Режим доступа: https://www.rcsb.org/structure/5N7V.

141. Schuller M., Correy G.J., Gahbauer S., Fearon D., Wu T., Díaz R.E., Young I.D., Carvalho Martins L., Smith D.H., Schulze-Gahmen U., Owens T.W., Deshpande I., Merz G.E., Thwin A.C., Biel J.T., Peters J.K., Moritz M., Herrera N., Kratochvil H.T., QCRG Structural Biology Consortium, Aimon A., Bennett J.M., Brandao Neto J., Cohen A.E., Dias A., Douangamath A., Dunnett L., Fedorov O., Ferla M.P., Fuchs M.R., Gorrie-Stone T.J., Holton J.M., Johnson M.G., Krojer T., Meigs G., Powell A.J., Rack J.G.M., Rangel V.L., Russi S., Skyner R.E., Smith C.A., Soares A.S., Wierman J.L., Zhu K., O'Brien P., Jura N., Ashworth A., Irwin J.J., Thompson M.C., Gestwicki J.E., von Delft F., Shoichet B.K., Fraser J.S., Ahel I. Fragment binding to the Nsp3 macrodomain of SARS-CoV-2 identified through crystallographic screening and computational docking // Science Advances. — 2021. — Т. 7. — № 16. — С. eabf8711

142. Newman J.A., Douangamath A., Yadzani S., Yosaatmadja Y., Aimon A., Brandão-Neto J., Dunnett L., Gorrie-Stone T., Skyner R., Fearon D., Schapira M., von Delft F., Gileadi O.

Structure, mechanism and crystallographic fragment screening of the SARS-CoV-2 NSP13 helicase // Nature Communications. — 2021. — Т. 12. — № 1. — С. 4848.

143. Douangamath A., Fearon D., Gehrtz P., Krojer T., Lukacik P., Owen C.D., Resnick E., Strain-Damerell C., Aimon A., Ábrányi-Balogh P., Brandão-Neto J., Carbery A., Davison G., Dias A., Downes T.D., Dunnett L., Fairhead M., Firth J.D., Jones S.P., Keeley A., Keserü G.M., Klein H.F., Martin M.P., Noble M.E.M., O'Brien P., Powell A., Reddi R.N., Skyner R., Snee M., Waring M.J., Wild C., London N., von Delft F., Walsh M.A. Crystallographic and electrophilic fragment screening of the SARS-CoV-2 main protease // Nature Communications. — 2020. — Т. 11. — № 1. — С. 5047.

144. PDB: 5R84 [Электронный ресурс]. Режим доступа: https://www.rcsb.org/structure/5R84.

145. Li Y., Liu Y., Wu J., Liu X., Wang L., Wang J., Yu J., Qi H., Qin L., Ding X., Ren F., Zhavoronkov A. Discovery of Potent, Selective, and Orally Bioavailable Small-Molecule Inhibitors of CDK8 for the Treatment of Cancer // Journal of Medicinal Chemistry. — 2023. — Т. 66. — № 8. — С. 5439–5452.

146. Zhu W., Liu X., Li Q., Gao F., Liu T., Chen X., Zhang M., Aliper A., Ren F., Ding X., Zhavoronkov A. Discovery of novel and selective SIK2 inhibitors by the application of AlphaFold structures and generative models // Bioorganic & Medicinal Chemistry. — 2023. — Т. 91. — С. 117414.

147. Wang Y., Wang C., Liu J., Sun D., Meng F., Zhang M., Aliper A., Ren F., Zhavoronkov A., Ding X. Discovery of 3-hydroxymethyl-azetidine derivatives as potent polymerase theta inhibitors // Bioorganic & Medicinal Chemistry. — 2024. — Т. 103. — С. 117662.

148. Xu J., Ding X., Fu Y., Meng Q., Wang L., Zhang M., Xu C., Chen S., Aliper A., Ren F., Zhavoronkov A., Ding X. Discovery of Novel and Potent Prolyl Hydroxylase Domain-Containing Protein (PHD) Inhibitors for The Treatment of Anemia // Journal of Medicinal Chemistry. — 2024. — Т. 67. — № 2. — С. 1393–1405.

149. Xu J., Qi H., Wang Z., Wang L., Steurer B., Cai X., Liu J., Aliper A., Zhang M., Ren F., Zhavoronkov A., Ding X. Discovery of a Novel and Potent Cyclin-Dependent Kinase 8/19 (CDK8/19) Inhibitor for the Treatment of Cancer // Journal of Medicinal Chemistry. — 2024. — Т. 67. — № 10. — С. 8161–8171.

150. Ren F., Ding X., Zheng M., Korzinkin M., Cai X., Zhu W., Mantsyzov A., Aliper A., Aladinskiy

V., Cao Z., Kong S., Long X., Man Liu B.H., Liu Y., Naumov V., Shneyderman A., Ozerov I.V., Wang J., Pun F.W., Polykovskiy D.A., Sun C., Levitt M., Aspuru-Guzik A., Zhavoronkov A. AlphaFold accelerates artificial intelligence powered drug discovery: efficient discovery of a novel CDK20 small molecule inhibitor // Chemical Science . — 2023. — Т. 14. — С. 1443–1452.

151. Jin Z., Du X., Xu Y., Deng Y., Liu M., Zhao Y., Zhang B., Li X., Zhang L., Peng C., Duan Y., Yu J., Wang L., Yang K., Liu F., Jiang R., Yang X., You T., Liu X., Yang X., Bai F., Liu H., Liu X., Guddat L.W., Xu W., Xiao G., Qin C., Shi Z., Jiang H., Rao Z., Yang H. Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors // Nature. — 2020. — Т. 582. — № 7811. — С. 289–293.

152. PDB: - 6W63 [Электронный ресурс]. Режим доступа: https://www.rcsb.org/structure/6W63.

153. Jacobs J., Grum-Tokars V., Zhou Y., Turlington M., Saldanha S.A., Chase P., Eggler A., Dawson E.S., Baez-Santos Y.M., Tomar S., Mielech A.M., Baker S.C., Lindsley C.W., Hodder P., Mesecar A., Stauffer S.R. Discovery, synthesis, and structure-based optimization of a series of N-(tert-butyl)-2-(N-arylamido)-2-(pyridin-3-yl) acetamides (ML188) as potent noncovalent small molecule inhibitors of the severe acute respiratory syndrome coronavirus (SARS-CoV) 3CL protease // Journal of Medicinal Chemistry. — 2013. — Т. 56. — № 2. — С. 534–546.

154. Патент US20170313685A1. Broad-spectrum non-covalent coronavirus protease inhibitors: опубл. 22.05.2018. / St. John S.E., Mesecar A.D.

155. Kombo D.C., LaMarche M.J. The logic of chemical optimization // Journal of Medicinal Chemistry. — 2025. — Т. 68. — № 11. — С. 11572–11585.

156. Shindo N., Fuchida H., Sato M., Watari K., Shibata T., Kuwata K., Miura C., Okamoto K., Hatsuyama Y., Tokunaga K., Sakamoto S., Morimoto S., Abe Y., Shiroishi M., Caaveiro J.M.M., Ueda T., Tamura T., Matsunaga N., Nakao T., Koyanagi S., Ohdo S., Yamaguchi Y., Hamachi I., Ono M., Ojida A. Selective and reversible modification of kinase cysteines with chlorofluoroacetamides // Nature Chemical Biology. — 2019. — Т. 15. — № 3. — С. 250–258.

157. Fu Y., Ding X., Zhang M., Feng C., Yan Z., Wang F., Xu J., Lin X., Ding X., Wang L., Fan Y., Li T., Yin Y., Liang X., Xu C., Chen S., Pulous F.E., Gennert D., Pun F.W., Kamya P., Ren F., Aliper A., Zhavoronkov A. Intestinal mucosal barrier repair and immune regulation with an AI-developed gut-restricted PHD inhibitor // Nature Biotechnology. — 2024. DOI: 10.1038/s41587-024-02503-w

158. Патент WO2023078238A1. SARS-CoV-2 inhibitors for treating coronavirus infections: опубл. 11.05.2023. / Ding X., Peng J., Ren F., Ding X., Zagribelnyy B., Ivanenkov Y.A.

159. Sun J., Sun D., Yang Q., Wang D., Peng J., Guo H., Ding X., Chen Z., Yuan B., Ivanenkov Y.A., Yuan J., Zagribelnyy B.A., He Y., Su J., Wang L., Tang J., Li Z., Li R., Li T., Hu X., Liang X., Zhu A., Wei P., Fan Y., Liu S., Zheng J., Guan X., Aliper A., Yang M., Bezrukov D.S., Xie Z., Terentiev V.A., Peng G., Polykovskiy D.A., Malyshev A.S., Malkov M.N., Zhu Q., Aspuru-Guzik A., Ding X., Cai X., Zhang M., Zhao J., Zhong N., Ren F., Chen X., Zhavoronkov A., Zhao J. A novel, covalent broad-spectrum inhibitor targeting human coronavirus Mpro // Nature Communications. — 2025. — Т. 16. — № 1. — С. 4546.

160. Бондарев Н., Загрибельный Б., Федорченко С.А., Иваненков Я.А., Палюлин В.А. Моделирование синтетической доступности потенциальных лекарственных веществ, содержащих пятичленные ароматические гетероциклы // Известия Академии наук. Серия химическая. — 2025. — Т. 74. — № 6. — С. 1687–1703. (Переводная версия: Bondarev N., Zagribelyy B., Fedorchenko S.A., Ivanenkov Ya. A., Palyulin V.A. Modeling of synthetic accessibility of potential drug molecules containing five-membered aromatic heterocycles // Russian Chemical Bulletin. — 2025. — Т. 74. — № 6. — С. 1687–1703.)

161. Ding H.X., Leverett C.A., Kyne R.E. Jr, Liu K.K.-C., Fink S.J., Flick A.C., O'Donnell C.J. Synthetic approaches to the 2013 new drugs // Bioorganic & Medicinal Chemistry. — 2015. — Т. 23. — № 9. — С. 1895–1922.

162. Flick A.C., Ding H.X., Leverett C.A., Kyne R.E. Jr, Liu K.K.-C., Fink S.J., O'Donnell C.J. Synthetic approaches to the 2014 new drugs // Bioorganic & Medicinal Chemistry. — 2016. — Т. 24. — № 9. — С. 1937–1980.

163. Ding H.X., Leverett C.A., Kyne R.E. Jr, Liu K.K.-C., Sakya S.M., Flick A.C., O'Donnell C.J. Synthetic approaches to the 2012 new drugs // Bioorganic & Medicinal Chemistry. — 2014. — Т. 22. — № 7. — С. 2005–2032.

164. Wang Y.-T., Yang P.-C., Zhang Y.-F., Sun J.-F. Synthesis and clinical application of new drugs approved by FDA in 2023 // European Journal of Medicinal Chemistry. — 2024. — Т. 265. — С. 116124.

165. Flick A.C., Leverett C.A., Ding H.X., McInturff E., Fink S.J., Mahapatra S., Carney D.W., Lindsey E.A., DeForest J.C., France S.P., Berritt S., Bigi-Botterill S.V., Gibson T.S., Liu Y., O'Donnell C.J. Synthetic Approaches to the New Drugs Approved during 2019 // Journal of

Medicinal Chemistry. — 2021. — Т. 64. — № 7. — С. 3604–3657.

166. WuXi Biologics [Электронный ресурс]. Режим доступа: https://www.wuxibiologics.com/.

167. Angene Chemical [Электронный ресурс]. Режим доступа: https://www.angenechemical.com/.

168. Crawford J.J., Johnson A.R., Misner D.L., Belmont L.D., Castanedo G., Choy R., Coraggio M., Dong L., Eigenbrot C., Erickson R., Ghilardi N., Hau J., Katewa A., Kohli P.B., Lee W., Lubach J.W., McKenzie B.S., Ortwine D.F., Schutt L., Tay S., Wei B., Reif K., Liu L., Wong H., Young W.B. Discovery of GDC-0853: A Potent, Selective, and Noncovalent Bruton's Tyrosine Kinase Inhibitor in Early Clinical Development // Journal of Medicinal Chemistry. — 2018. — Т. 61. — № 6. — С. 2227–2245.

169. Flick A.C., Leverett C.A., Ding H.X., McInturff E.L., Fink S.J., Mahapatra S., Carney D.W., Lindsey E.A., DeForest J.C., France S.P., Berritt S., Bigi-Botterill S.V., Gibson T.S., Watson R.B., Liu Y., O'Donnell C.J. Synthetic Approaches to the New Drugs Approved During 2020 // Journal of Medicinal Chemistry. — 2022. — Т. 65. — № 14. — С. 9607–9661.

170. McCann S.D., Reichert E.C., Arrechea P.L., Buchwald S.L. Development of an Aryl Amination Catalyst with Broad Scope Guided by Consideration of Catalyst Stability // Journal of the American Chemical Society. — 2020. — Т. 142. — № 35. — С. 15027–15037.

171. Hong Y., Senanayake C.H., Xiang T., Vandenbossche C.P., Tanoury G.J., Bakale R.P., Wald S.A. Remarkably selective palladium-catalyzed amination process: Rapid assembly of multiamino based structures // Tetrahedron Letters. — 1998. — Т. 39. — № 20. — С. 3121–3124.

172. Cabello-Sanchez N., Jean L., Maddaluno J., Lasne M.-C., Rouden J. Palladium-mediated N-arylation of heterocyclic diamines: insights into the origin of an unusual chemoselectivity // The Journal of Organic Chemistry. — 2007. — Т. 72. — № 6. — С. 2030–2039.

173. Segler M.H.S., Kogej T., Tyrchan C., Waller M.P. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks // ACS Central Science. — 2018. — Т. 4. — № 1. — С. 120–131.

174. Skalic M., Jiménez J., Sabbadin D., De Fabritiis G. Shape-Based Generative Modeling for de Novo Drug Design // Journal of Chemical Information and Modeling. — 2019. — Т. 59. — № 3. — С. 1205–1214.

175. Wei L., Wen W., Rao L., Huang Y., Lei M., Liu K., Hu S., Song R., Ren Y., Wan J. Cov_FB3D: A De Novo Covalent Drug Design Protocol Integrating the BA-SAMP Strategy and Machine-Learning-Based Synthetic Tractability Evaluation // Journal of Chemical Information and Modeling. — 2020. — Т. 60. — № 9. — С. 4388–4402.

176. Bung N., Krishnan S.R., Bulusu G., Roy A. De novo design of new chemical entities for SARS-CoV-2 using artificial intelligence // Future Medicinal Chemistry. — 2021. — Т. 13. — № 6. — С. 575–585.

177. Krishnan S.R., Bung N., Bulusu G., Roy A. Accelerating De Novo Drug Design against Novel Proteins Using Deep Learning // Journal of Chemical Information and Modeling. — 2021. — Т. 61. — № 2. — С. 621–630.

178. Srinivas R., Verma N., Kraka E., Larson E.C. Deep Learning-Based Ligand Design Using Shared Latent Implicit Fingerprints from Collaborative Filtering // Journal of Chemical Information and Modeling. — 2021. — Т. 61. — № 5. — С. 2159–2174.

179. Bilsland A.E., McAulay K., West R., Pugliese A., Bower J. Automated Generation of Novel Fragments Using Screening Data, a Dual SMILES Autoencoder, Transfer Learning and Syntax Correction // Journal of Chemical Information and Modeling. — 2021. — Т. 61. — № 6. — С. 2547–2559.

180. Yang L., Yang G., Bing Z., Tian Y., Niu Y., Huang L., Yang L. Transformer-Based Generative Model Accelerating the Development of Novel BRAF Inhibitors // ACS Omega. — 2021. — Т. 6. — № 49. — С. 33864–33873.

181. Srinivasan S., Batra R., Chan H., Kamath G., Cherukara M.J., Sankaranarayanan S.K.R.S. Artificial Intelligence-Guided De Novo Molecular Design Targeting COVID-19 // ACS Omega. — 2021. — Т. 6. — № 19. — С. 12557–12566.

182. Hu L., Yang Y., Zheng S., Xu J., Ran T., Chen H. Kinase Inhibitor Scaffold Hopping with Deep Learning Approaches // Journal of Chemical Information and Modeling. — 2021. — Т. 61. — № 10. — С. 4900–4912.

183. Krishnan S.R., Bung N., Vangala S.R., Srinivasan R., Bulusu G., Roy A. De Novo Structure-Based Drug Design Using Deep Learning // Journal of Chemical Information and Modeling. — 2022. — Т. 62. — № 21. — С. 5100–5109.

184. Creanza T.M., Lamanna G., Delre P., Contino M., Corriero N., Saviano M., Mangiatordi G.F.,

Ancona N. DeLA-Drug: A Deep Learning Algorithm for Automated Design of Druglike Analogues // Journal of Chemical Information and Modeling. — 2022. — Т. 62. — № 6. — С. 1411–1424.

185. Bung N., Krishnan S.R., Roy A. An In Silico Explainable Multiparameter Optimization Approach for De Novo Drug Design against Proteins from the Central Nervous System // Journal of Chemical Information and Modeling. — 2022. — Т. 62. — № 11. — С. 2685–2695.

186. Zabolotna Y., Volochnyuk D.M., Ryabukhin S.V., Gavrylenko K., Horvath D., Klimchuk O., Oksiuta O., Marcou G., Varnek A. SynthI: A New Open-Source Tool for Synthon-Based Library Design // Journal of Chemical Information and Modeling. — 2022. — Т. 62. — № 9. — С. 2151–2163.

187. Krishnan S.R., Bung N., Padhi S., Bulusu G., Misra P., Pal M., Oruganti S., Srinivasan R., Roy A. De novo design of anti-tuberculosis agents using a structure-based deep learning method // Journal of Molecular Graphics & Modelling. — 2023. — Т. 118. — С. 108361.

188. Ursu O., Glick M., Oprea T. Novel drug targets in 2018 // Nature Reviews. Drug Discovery. — 2019. — Т. 18. — № 5. — С. 328–328.

189. Avram S., Halip L., Curpan R., Oprea T.I. Novel drug targets in 2020 // Nature Reviews. Drug Discovery. — 2021. — Т. 20. — № 5. — С. 333.

190. Avram S., Halip L., Curpan R., Oprea T.I. Novel drug targets in 2021 // Nature Reviews. Drug Discovery. — 2022. — Т. 21. — № 5. — С. 328.

191. Avram S., Halip L., Curpan R., Oprea T.I. Novel drug targets in 2022 // Nature Reviews. Drug Discovery. — 2023. — Т. 22. — № 6. — С. 437.

192. Ran X., Gestwicki J.E. Inhibitors of protein-protein interactions (PPIs): an analysis of scaffold choices and buried surface area // Current Opinion in Chemical Biology. — 2018. — Т. 44. — С. 75–86.

193. Macalino S.J.Y., Basith S., Clavio N.A.B., Chang H., Kang S., Choi S. Evolution of in silico strategies for protein-protein interaction drug discovery // Molecules (Basel, Switzerland). — 2018. — Т. 23. — № 8. — С. 1963.

194. Scott D.E., Bayly A.R., Abell C., Skidmore J. Small molecules, big targets: drug discovery faces the protein-protein interaction challenge // Nature Reviews. Drug Discovery. — 2016. — Т. 15.

— № 8. — С. 533–550.

195. Soga S., Shirai H., Kobori M., Hirayama N. Use of amino acid composition to predict ligand-binding sites // Journal of Chemical Information and Modeling. — 2007. — Т. 47. — № 2. — С. 400–406.

196. Ivanenkov Y.A., Zagribelnyy B.A., Aladinskiy V.A. Are we opening the door to a New Era of medicinal chemistry or being collapsed to a chemical singularity? // Journal of Medicinal Chemistry. — 2019. — Т. 62. — № 22. — С. 10026–10043.

Приложение А

Таблица А1 Фрагмент иерархической библиотеки SMARTS-подструктур пятичленных ароматических гетероциклов (перенос SMARTS на новую строку обозначен символом ↲)

Слой	ID	Описание	SMARTS
0	0	пятичленный ароматический гетероцикл	[a:1]:1:a:a:a:a1
1	1	пятичленный ароматический гетероцикл с одним гетероатомом	[a:1]:1:[#6](*):[#6](*):[#6](*):[#6](*)1
1	2	1,2-азол	[a:1]:1:[#7+0]:[#6](*):[#6](*):[#6](*)1
1	3	1,3-азол	[a:1]:1:[#6](*):[#7+0]:[#6](*):[#6](*)1
1	4	1,2,3-диазол	[a:1]:1:[#7+0]:[#7+0]:[#6](*):[#6](*)1
1	5	1,2,4-диазол	[a:1]:1:[#7+0]:[#6](*):[#7+0]:[#6](*)1
1	6	1,2,5-диазол	[a:1]:1:[#7+0]:[#6](*):[#6](*):[#7+0]1
1	7	1,3,4-диазол	[a:1]:1:[#7+0]:[#6](*):[#6](*):[#7+0]1
2	1.1	пиррол	[#7+0](*):1:[#6](*):[#6](*):[#6](*):[#6](*)1
2	1.2	фуран	[#8]:1:[#6](*):[#6](*):[#6](*):[#6](*)1
2	1.3	тиофен	[#16]:1:[#6](*):[#6](*):[#6](*):[#6](*)1
2	2.1	пиразол	[#7+0](*):1:[#7+0]:[#6](*):[#6](*):[#6](*)1
2	2.2	изоксазол	[#8]:1:[#7+0]:[#6](*):[#6](*):[#6](*)1
2	2.3	изотиазол	[#16]:1:[#7+0]:[#6](*):[#6](*):[#6](*)1
2	…	…	…
3	1.1.1	1,2-дизамещенный пиррол	[#7+0](*):1:[#6](*):[#6]([#1]):[#6]([#1]):[#6]([#1])1
3	1.1.2	1,3-дизамещенный пиррол	[#7+0](*):1:[#6]([#1]):[#6](*):[#6]([#1]):[#6]([#1])1
3	1.1.3	2,3-дизамещенный пиррол	[#7+0]([#1]):1:[#6](*):[#6](*):[#6]([#1]):[#6]([#1])1
3	1.1.4	2,4-дизамещенный пиррол	[#7+0]([#1]):1:[#6](*):[#6]([#1]):[#6](*):[#6]([#1])1
3	…	…	…
3	1.1.7	1,2,3-тризамещенный пиррол	[#7+0](*):1:[#6](*):[#6](*):[#6]([#1]):[#6]([#1])1

Продолжение на следующей странице

Продолжение таблицы А1

Слой	ID	Описание	SMARTS
3	…	…	…
3	1.2.1	2,3-дизамещенный фуран	[#8]:1:[#6](*):[#6](*):[#6]([#1]):[#6]([#1])1
3	1.2.2	2,4-дизамещенный фуран	[#8]:1:[#6](*):[#6]([#1]):[#6](*):[#6]([#1])1
3	…	…	…
3	2.1.1	1,3-дизамещенный пиразол	[#7+0](*):1:[#7+0]:[#6](*):[#6]([#1]):[#6]([#1])1
3	2.1.2	1,4-дизамещенный пиразол	[#7+0](*):1:[#7+0]:[#6]([#1]):[#6](*):[#6]([#1])1
3	…	…	…
4	1.1.1.1	1,2-C1-дизамещенный пиррол	[#7+0](!=!@[CH3,CH2]):1:[#6](!=!@[CH3,CH2]):[#6]([#1]):[#6]([#1]):[#6]([#1])1
4	1.1.1.2	1-C1-2-C2-замещенный пиррол	[#7+0](!=!@[CH3,CH2]):1:[#6](!=!@[CH,CH0]):[#6]([#1]):[#6]([#1]):[#6]([#1])1
4	1.1.1.3	1-C1-2-C3-замещенный пиррол	[#7+0](!=!@[CH3,CH2]):1:[#6](!=!:[c]):[#6]([#1]):[#6]([#1]):[#6]([#1])1
4	…	…	…
4	1.1.2.1	1,3-C1-дизамещенный пиррол	[#7+0](*):1:[#6]([#1]):[#6](*):[#6]([#1]):[#6]([#1])1
4	1.1.2.2	1-C1-3-C2-замещенный пиррол	[#7+0](*):1:[#6]([#1]):[#6](*):[#6]([#1]):[#6]([#1])1
4	…	…	…
4	1.1.7.1	1,2,3-C1-тризамещенный пиррол	[#7+0](!=!@[CH3,CH2]):1:[#6]([#1]):[#6](!=!@[CH3,CH2]):[#6]([#1]):[#6]([#1])1
4	1.1.7.2	1-C1-2-C1-3-C2-замещенный пиррол	[#7+0](!=!@[CH3,CH2]):1:[#6]([#1]):[#6](!=!@[CH,CH0]):[#6]([#1]):[#6]([#1])1
4	1.1.7.3	1-C1-2-C1-3-C3-замещенный пиррол	[#7+0](!=!@[CH3,CH2]):1:[#6]([#1]):[#6](!=!:[c]):[#6]([#1]):[#6]([#1])1
4	…	…	…
4	2.1.1.1	1,3-C1-дизамещенный пиразол	[#7+0](!=!@[CH3,CH2]):1:[#7+0]:[#6](!=!@[CH3,CH2]):[#6]([#1]):[#6]([#1])1
4	2.1.1.2	1-C1-3-C2-замещенный пиразол	[#7+0](!=!@[CH3,CH2]):1:[#7+0]:[#6](!=!@[CH,CH0]):[#6]([#1]):[#6]([#1])1
4	…	…	…
4	2.1.2.1	1,4-C1-дизамещенный пиразол	[#7+0](!=!@[CH3,CH2]):1:[#7+0]:[#6]([#1]):[#6](!=!@[CH3,CH2]):[#6]([#1])1
4	2.1.2.2	1-C1-4-C2-замещенный пиразол	[#7+0](!=!@[CH3,CH2]):1:[#7+0]:[#6]([#1]):[#6](!=!@[CH,CH0]):[#6]([#1])1
4	…	…	…