Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение
высшего образования
«Уральский федеральный университет
имени первого Президента России Б.Н. Ельцина»
Институт радиоэлектроники и информационных технологий – РТФ
Школа профессионального и академического образования

ДОПУСТИТЬ К ЗАЩИТЕ ПЕРЕД ГЭК

Директор ШПиАО
(подпись) Д.В. Денисов (Ф.И.О.)
« 03 » июня 2024 г.

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

ОЦЕНКА ВЛИЯНИЯ ДАННЫХ НА КАЧЕСТВО РАБОТЫ МОДЕЛИ ПРЕДСКАЗАНИЯ СХЕМ СИНТЕЗА ОРГАНИЧЕСКИХ МОЛЕКУЛ

Научный руководитель: Долганов Антон Юрьевич
к.т.н., доцент
подпись

Нормоконтролер: Огуренко Егор Владимирович
подпись

Студент группы: РИМ-220963 Голубев Артём Алексеевич
подпись

Екатеринбург
2024

Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение высшего образования
«Уральский федеральный университет имени первого Президента России Б.Н. Ельцина»

Институт радиоэлектроники и информационных технологий – РТФ
Школа профессионального и академического образования
Направление подготовки 09.04.01 Информатика и вычислительная техника
Образовательная программа 09.04.01/33.03 Инженерия машинного обучения

ЗАДАНИЕ

на выполнение выпускной квалификационной работы

студента *Голубева Артёма Алексеевича* **группы** *РИМ-220963*
(фамилия, имя, отчество)

1. Тема выпускной квалификационной работы
Оценка влияния данных на качество работы модели предсказания схем синтеза органических молекул
Утверждена распоряжением по институту от «4» декабря 2023 г. № 33.02-05/298

2. Научный руководитель
Долганов Антон Юрьевич, кандидат технических наук, доцент
(Ф.И.О., должность, ученая степень, ученое звание)

3. Исходные данные к работе
Материалы, полученные в ходе преддипломной практики

4. Перечень демонстрационных материалов
Презентация в MS PowerPoint

5. Календарный план

№ п/п	Наименование этапов выполнения работы	Срок выполнения этапов работы	Отметка о выполнении
1.	*1 раздел (глава)*	до 23.03.2024 г.	
2.	*2 раздел (глава)*	до 29.04.2024 г.	
3.	*3–4 раздел (глава)*	до 20.05.2024 г.	
4.	*ВКР в целом*	до 24.05.2024 г.	

Научный руководитель Долганов Антон Юрьевич _____ (подпись)
Ф.И.О.

Студент задание принял к исполнению 12.02.24 _____ (подпись)
дата

6. Допустить Голубева Артёма Алексеевича **к защите выпускной квалификационной работы в экзаменационной комиссии**

Директор ШПиАО _____ Д.В. Денисов
(подпись) Ф.И.О.

РЕФЕРАТ

Выпускная квалификационная работа магистра 81 с., 43 рис., 2 табл., 66 источников.

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ, РЕТРОСИНТЕТИЧЕСКИЙ АНАЛИЗ, ИНЖИНИРИНГ ДАННЫХ, ОЦЕНКА КАЧЕСТВА ПРЕДСКАЗАНИЙ

Цель работы – оценка влияния данных на качество работы модели предсказания схем синтеза органических молекул и интеграция данной модели в собственный вычислительный веб-сервис.

Объект исследования – предсказание схем синтеза органических молекул с помощью ИИ-моделей.

Методы исследования: анализ литературы по теме исследования; изучение документации исследуемого ретросинтетического инструмента; тестирование инструмента с различными параметрами анализа; сбор и обработка данных для запуска анализа; обобщение полученных результатов и их сравнение; разработка программного обеспечения для последующего внедрения исследуемого инструмента в собственный вычислительный сервис.

Результаты работы: в ходе тестирования инструмента были подобраны оптимальные параметры запуска; исследуемый ретросинтетический инструмент был внедрен в собственный вычислительный веб-сервис.

СОДЕРЖАНИЕ

РЕФЕРАТ 3
СОДЕРЖАНИЕ 4
ВВЕДЕНИЕ 6
1 Литературный обзор 9
1.1 Ключевые подходы к автоматизации ретросинтеза 9
1.2 Ключевые понятия ретросинтеза 12
1.3 Подходы к ретросинтезу на основе МО и ИИ 17
1.4 Вывод 28
2 Методическая часть 31
2.1 AiZynthFinder 31
2.1.1 Архитектура 31
2.1.2 Ввод и вывод инструмента 35
2.1.3 Ключевые параметры запуска 37
2.2 Методология оценки качества ретросинтеза 38
2.3 Вывод 41
3 Обсуждение результатов 42
3.1 Запуск анализа в стандартных условиях 42
3.2 Влияние простых параметров запуска 44
3.2.1 Время анализа 44
3.2.2 Глубина анализа 45
3.2.3 Выводы по простым параметрам 48
3.3 Влияние стоков 49
3.3.1 Сток с in-house данными 50
3.3.2 Стоки из открытых баз данных 52
3.3.3 Использование комбинаций стоков 56

3.3.4 Выводы по стокам 59
3.4 Влияние стратегий расширения 60
3.5 Валидация результатов 61
3.6 Вывод 64
4 Интеграция инструмента 67
4.1 Архитектура 67
4.2 Пользовательский интерфейс 68
4.3 Технические особенности 70
4.4 Вывод 71
ЗАКЛЮЧЕНИЕ 72
СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 75

ВВЕДЕНИЕ

Не будет преувеличением сказать, что в последние годы алгоритмы машинного обучения (МО) и искусственного интеллекта (ИИ) получили широкое распространение во всех сферах окружающей нас действительности [1]. Химия и, в частности, органический синтез не стали исключением. Одним из направлений применения технологий ИИ и МО в органической химии является ретросинтетический анализ, позволяющий декомпозировать сложные органические молекулы на более простые компоненты (билдинг-блоки, building-blocks) для составления подходящих схем синтеза.

Создание эффективных и реализуемых синтетических схем требует от химика-синтетика большое количество усилий. Во-первых, для выбора наиболее подходящих путей разбиения молекул необходимо иметь большой опыт и «насмотренность» в органической химии. Во-вторых, составление схем ретросинтеза зачастую требует довольно трудоемкого и долгого анализа литературных источников на предмет практической реализуемости тех или иных химических трансформаций. В-третьих, немало времени может уйти на поиск подходящих реактивов у поставщиков химической продукции и в собственном каталоге веществ. Также стоит отметить, что время- и трудозатраты увеличиваются кратно количеству молекул, для которых необходимо составить схемы синтеза. Все эти аспекты делают автоматизацию данного процесса актуальной задачей.

Ключевыми преимуществами автоматизированного ретросинтетического анализа являются упрощение и, главное, заметное ускорение поиска путей синтеза новых соединений, что критически важно для фармацевтических компаний при разработке лекарственных препаратов на основе малых молекул. Кроме того, благодаря большим обучающим выборкам, основанным на химических базах данных с миллионами реакций, такие инструменты способны предлагать не самые очевидные, но при этом

действенные пути синтеза, оптимизируя их с точки зрения доступности реактивов и других параметров.

Среди инструментов, предназначенных для автоматизации ретросинтетического анализа, особенно выделяется AiZynthFinder – программа с открытым исходным кодом, использующая нейросетевые модели для предсказания реакционных путей. Ключевым преимуществом данного инструмента является возможность его тонкой настройки под пользовательские нужды и, как следствие, удобство внедрения в собственные сервисы.

Стоит отметить, что качество работы AiZynthFinder напрямую зависит от данных, которые используются для запуска предсказаний. Так, при запуске инструмента со стандартной настройкой «из коробки» алгоритм чаще всего выдает достаточно сомнительные результаты с точки зрения применимости предлагаемых превращений. Тем не менее, за счет гибкой настройки параметров анализа и инжиниринга входных данных можно добиться значительного улучшения качества предсказаний и получения приемлемых химических схем.

Таким образом, **объект исследования** – предсказание схем синтеза с помощью ИИ-моделей; **предмет исследования** – влияние данных на качество работы ретросинтетического инструмента.

Целью работы является оценка влияния качества данных на результаты ретросинтетического анализа, реализуемого с помощью инструмента AiZynthFinder, и интеграция данного инструмента в собственный вычислительный веб-сервис фармацевтической компании Biocad.

Для достижения поставленной цели необходимо решить следующие **задачи**:

1) Провести анализ литературы по методам автоматизации ретросинтетического анализа;

2) Описать архитектуру AiZynthFinder и методологию тестирования инструмента;

3) Подготовить подходящие данные, провести тестирование инструмента и описать оптимальные параметры запуска ретросинтетического анализа;

4) Внедрить модель AiZynthFinder в собственный вычислительный веб-сервис компании Biocad.

При выполнении работы использовались следующие **методы**: анализ литературы, изучение документации, сбор и обработка данных, тестирование инструмента, обобщение и сравнение полученных результатов, разработка программного обеспечения, интеграция модели в собственный вычислительный сервис.

Научная новизна. В рамках данной работы впервые было проведено исследование влияния простых параметров и стоков на качество работы ретросинтетической модели AiZynthFinder в контексте построения синтетических схем.

Публикации. Результаты работы были опубликованы в сборнике тезисов Всероссийской научной студенческой конференции «ИНТЕР – Информационные технологии и радиоэлектроника 2024» (ISBN: 978-5-91256-646-2).

Для выполнения работы были использованы научные публикации по теме исследования, техническая документация использованных программных пакетов, материалы преддипломной практики, бизнес-требования заказчика.

Результатом работы стал подбор оптимальных параметров запуска ретросинтетического анализа для модели AiZynthFinder, а также интеграция данной модели в вычислительный сервис компании Biocad.

1 Литературный обзор

1.1 Ключевые подходы к автоматизации ретросинтеза

Ретросинтетический анализ – это подход к планированию схем синтеза органических молекул, заключающийся в последовательной декомпозиции целевой молекулы на всё более простые компоненты (прекурсоры) до тех пор, пока предшественниками не окажутся коммерчески доступные простые вещества [2]. С учетом итеративной сущности ретросинтетического анализа данный подход отлично формализуется и, как следствие, автоматизируется.

Первопроходцами в автоматизации ретросинтеза стали американские ученые Кори и Уипке, представившие в 1969 году свой алгоритм для автоматического ретросинтеза и основанный на нем инструмент OCSS (Organic Chemical Simulation of Synthesis) [3]. В рамках работы программы введенная пользователем структура анализируется на предмет особенностей химического строения, а затем разбивается на составные части на основании заранее закодированных правил декомпозиции и трансформации молекул (рисунок 1).

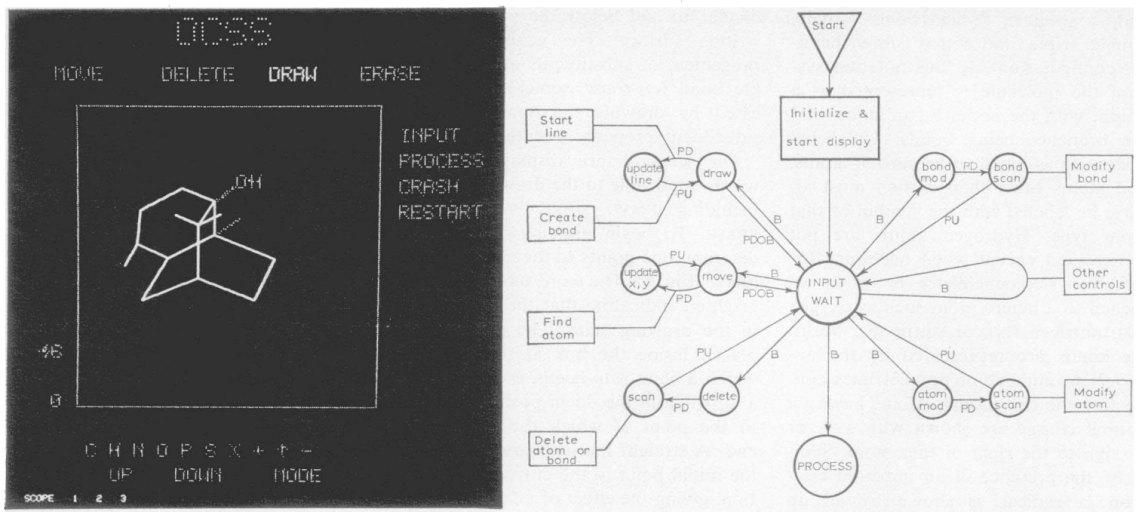

Рисунок 1 – Интерфейс и схема работы программы OCSS [3]

Данная программа стала основоположником первого подхода для автоматизации ретросинтеза: так называемых «экспертных систем на основе правил» (rule-based expert systems), активно развивавшихся в течение нескольких десятков лет [4]. Ключевой особенностью архитектуры таких алгоритмов является использование вручную закодированных реакционных правил (ретросинтетических шаблонов), согласно которым происходят трансформации молекул. Каждый такой шаблон представляет собой молекулярный подграф химической реакции и иллюстрирует изменения в связях между ключевыми атомами в ходе превращения молекул-реагентов (рисунок 2). При этом правила, по которым данные шаблоны применяются к целевой и последующим молекулам для построения ретросинтетического дерева, также задаются вручную.

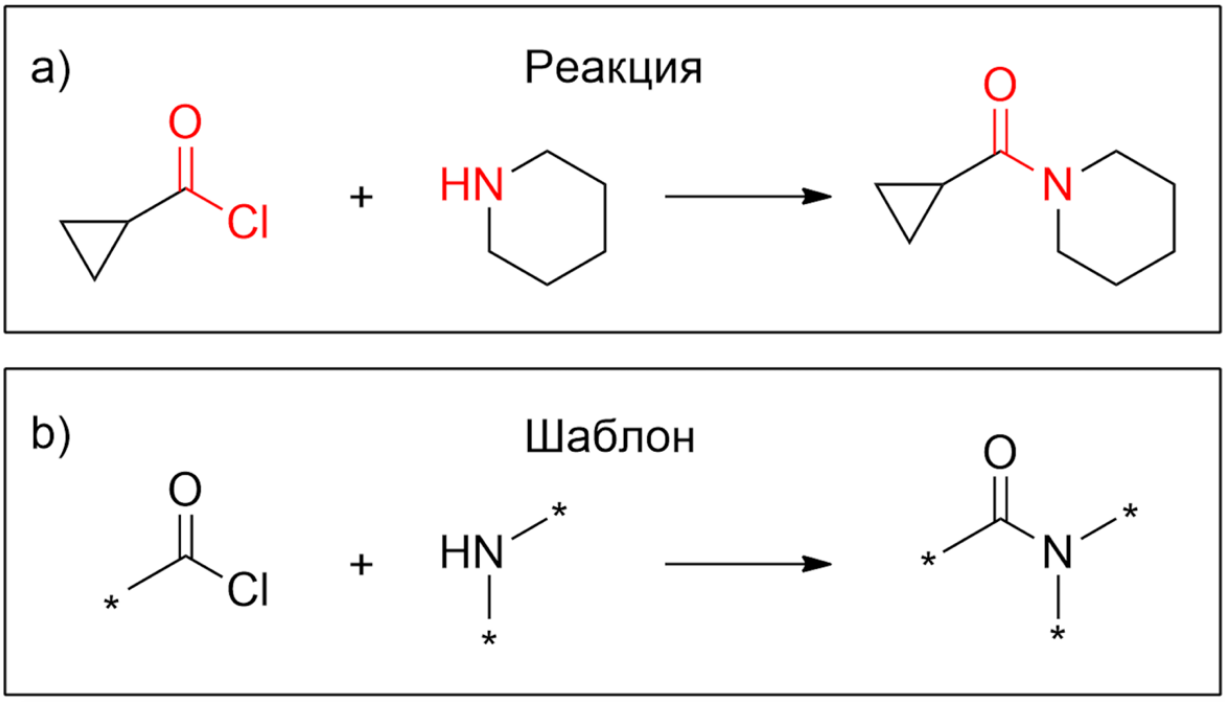

Рисунок 2 – Иллюстрация химической реакции и ретросинтетического шаблона: a) отмеченные красным части относятся к реакционному центру; b) отмеченные звездочкой части указывают на место крепления к молекуле за пределами шаблона

Несмотря на то, что «экспертные системы» зачастую могут похвастаться довольно надежными предсказаниями за счёт введения проверенных человеком правил [5], они имеют ряд существенных недостатков. Ключевой проблемой такого подхода является невозможность предсказания ретросинтетических путей для новых молекул из-за отсутствия в «базах знаний» новых реакционных шаблонов и реагентов [6]. Другим ограничением «экспертных систем» часто является их быстродействие [7]: из-за постоянного перебора шаблонов такие алгоритмы требуют на свою работу больших вычислительных затрат. Кроме того, кодирование реакционных правил человеком – это крайне трудоёмкий и медленный процесс [8].

В последние годы благодаря значительному увеличению вычислительных мощностей и накоплению большого количества данных стремительно развивается второй подход для автоматизации ретросинтеза: а именно использование моделей на основе искусственного интеллекта и машинного обучения, позволяющих преодолеть недостатки «экспертных систем» [9]. Главным преимуществом ИИ-моделей в задаче ретросинтетического планирования является их обучение на обширных экспериментальных датасетах, включающих в себя огромное количество самых разнообразных реакционных шаблонов. Во множестве работ было показано, что такой подход позволяет значительно превзойти результаты «экспертных систем» [10; 11]. Рассмотрению существующих ретросинтетических подходов на основе использования искусственного интеллекта и машинного обучения будет посвящен подраздел 1.3 литературного обзора.

Также заслуживает внимания третий подход к построению ретросинтетических схем: планировщики синтеза. Данный подход не основан на использовании шаблонов реакций как в ранее рассмотренных примерах. Вместо этого планировщики синтеза производят поиск подходящих реакций по базам данных, таким как SciFinder [12] и Reaxys [13]. Стоит отметить, что

сейчас в чистом виде планировщики синтеза почти не представлены: ретросинтетические инструменты на их основе в последние годы стали гибридными и теперь совмещают поиск по базе данных с использованием ИИ-моделей или других решений. Примерами таких комбинированных инструментов являются Reaxys Predictive Retrosynthesis [14], SciFinder Synthesis Planning [15], SYNTHIA Retrosynthesis Software [16]. Главным недостатком подобных решений является то, что распространяются они на коммерческой основе и зачастую имеют ряд региональных ограничений – это делает данные инструменты недоступными для широкого использования и тонкой настройки под свои нужды.

1.2 Ключевые понятия ретросинтеза

Прежде чем приступать к рассмотрению ретросинтетических инструментов на основе искусственного интеллекта и машинного обучения, стоит обозначить и в некоторых случаях формализовать ряд сопутствующих понятий [9–11].

1. Молекула

Для молекул существует множество способов формализации, здесь будут описаны ключевые из них, используемые в данной работе (рисунок 3):

– Графы. Молекула, состоящая из n атомов, описывается ненаправленным графом $G = (A, X)$, где A – множество атомов (вершин), а X – множество химических связей (ребер) [11];

– SMILES – строка упрощенной системы ввода молекулярных данных (Simplified Molecular Input Line Entry System) [17], в которой каждый символ обозначает определенный структурный элемент: атом, тип химической связи, ветвление и т.д. Для одной молекулы может быть несколько равнозначных SMILES-представлений, т.е. SMILES могут быть не уникальными;

– InChiKey – международный текстовый химический идентификатор (International Chemical Identifier) [18], уникальная хеш-строка для стандартного обозначения молекулы.

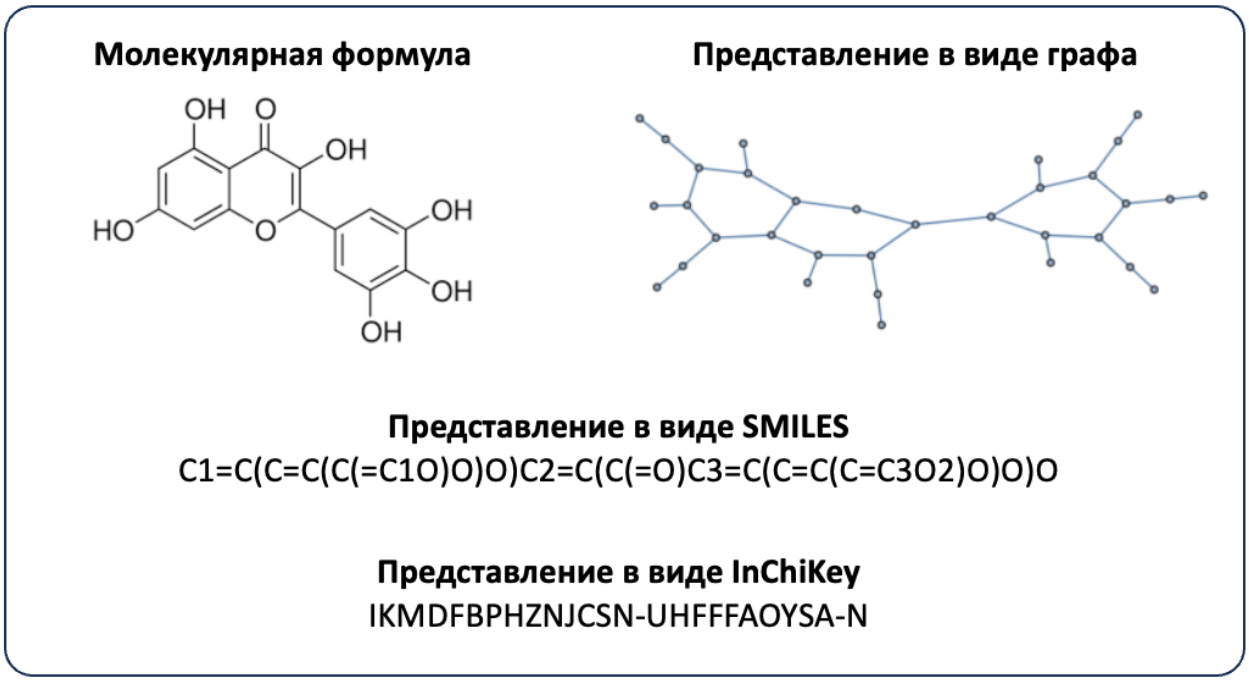

Рисунок 3 – Виды представления молекулы

2. Реакция

Химическая реакция – это преобразование одного набора молекул в другой набор вида $\{R_{i=1}^{n_r}\} \rightarrow \{P_{j=1}^{n_p}\}$, где Ri – i-тая молекула-реагент (исходная молекула), Pj – j-тая молекула-продукт (конечная молекула), а n_r и n_p – число молекул-реактантов и число молекул-продуктов соответственно (см. рисунок 2-а).

3. Одностадийная ретросинтетическая реакция

Одностадийная ретросинтетическая реакция – это инвертированная химическая реакция $P \implies \{R_{i=1}^{n_r}\}$, предсказывающая набор молекул-реагентов $\{R_{i=1}^{n_r}\}$, которые могли бы привести к получению желаемого продукта P (рисунок 4). Стрелка вида " $\implies$ " указывает на ретросинтетический характер

превращения. Стоит отметить, что в рамках данной работы рассматриваются только одностадийные ретросинтетические реакции с единственным продуктом P (однокомпонентные реакции), а реакции с большим количеством продуктов (многокомпонентные реакции) рассматриваются как несколько однокомпонентных реакций.

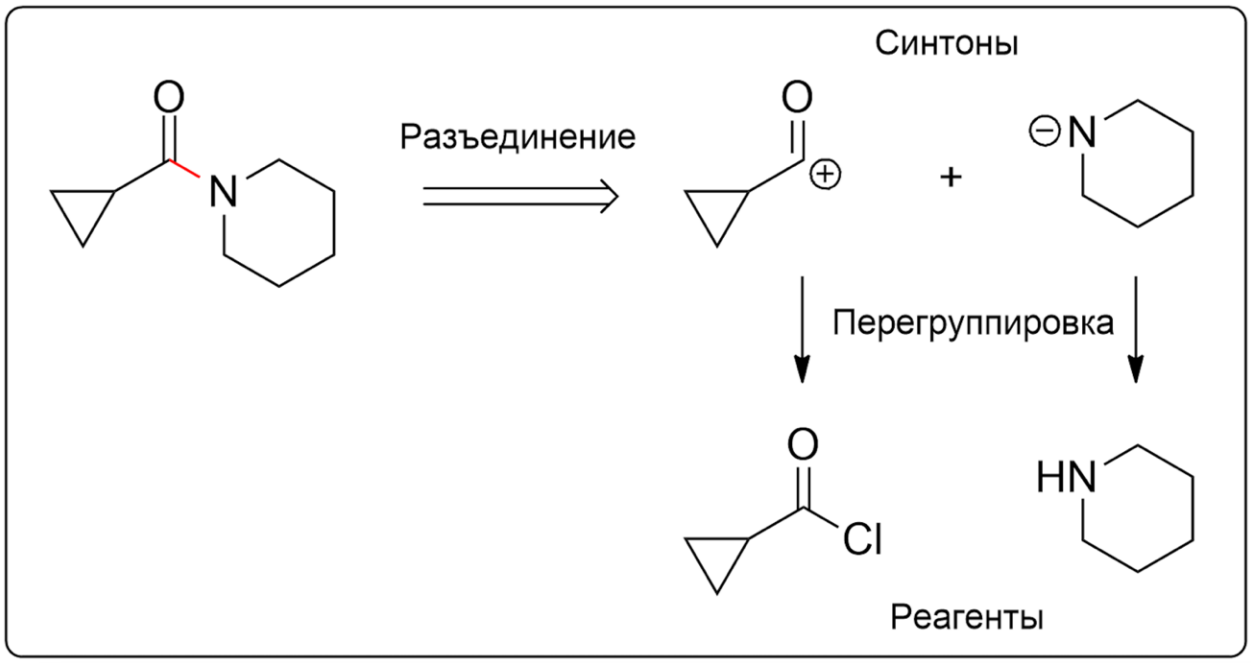

Рисунок 4 – Одностадийная ретросинтетическая реакция

Основной метрикой оценки качества одностадийного ретросинтетического предсказания при обучении моделей является точность (Accuracy) – доля правильных ответов среди всех предсказаний алгоритма для тестового набора молекул. Правильным ответом считается верно предзаказанное правило реакции. Обычно в дополнение к Accuracy считаются также метрики TopNAcc (Top10Acc, Top50Acc): предсказания модели сортируются в порядке уменьшения вероятности ответа, затем среди полученного сортированного списка выделяется N первых предсказаний, для которых определяется наличие верно предсказанного реакционного правила.

4. Сопоставление атомов

Предсказание как химической, так и ретросинтетической реакции осуществляется в соответствии с принципом сопоставления атомов (своеобразный закон сохранения атомов). Этот принцип гласит, что каждому атому в реагентах соответствует ровно один атом в продуктах и наоборот. Данное фундаментальное соотношение "один к одному" физически ограничивает пространство реакций и определяет, что химические реакции в основном связаны с разрывом и образованием химических связей [19].

5. Реакционный центр

Реакционный центр – для химической реакции это подмножество атомных пар $C = \{(a_i, a_j)\} \subseteq A \times A$, которые меняют типы связей во время химической реакции с образованием продукта P. В случае одностадийной ретросинтетической реакции реакционный центр – это подмножество существующих химических связей $C = \{x_i\} \subseteq X$ в продукте P, которые могут быть изменены для получения молекул-реагентов. С точки зрения химии реакционный центр включает атомы и связи, которые непосредственно участвуют в образовании связей и перегруппировке электронов в реакции [20]. Определение реакционного центра является важным аспектом как химических, так и ретросинтетических реакций.

6. Синтон

При разъединении связей в реакционном центре продукта P получается набор синтонов $\{S_i\}_{i=1}^{n_r}$. Синтон S_i – это молекулярный подграф реагента R_i, не обязательно являющийся подходящей молекулой. По сути, синтон является виртуальной частицей с условным зарядом ($\oplus$ или $\ominus$, рисунок 4), введенной для удобства подбора реальных реагентов в рамках ретросинтетического анализа [2].

7. Шаблон

Ретросинтетический шаблон T – это отображение следующего вида:

$$T := p^T \Longrightarrow \{r_i^T\}_{i=1}^{n_r}$$

Здесь p^T – это молекулярный подграф продукта P (данный подграф может рассматриваться как реакционный центр), а r_i^T – это молекулярный подграф i-ого реагента. Более наглядно ретросинтетический шаблон представлен на рисунке 2-b.

8. Многостадийное ретросинтетическое планирование

При наличии целевой молекулы P (target-molecule) многостадийное ретросинтетическое планирование направлено на прогнозирование последовательности реакций $\{r_d\}_{d=1}^{d_{max}}$, где d_{max} – это длина ретросинтетического пути (максимальное число стадий в схеме), а r_d – это одностадийная ретросинтетическая реакция. Данная последовательность продолжается до тех пор, пока все требуемые для целевой молекулы P реагенты не будут соответствовать коммерчески доступным соединениям. Если последовательность достигает максимальной длины схемы d_{max}, но при этом в какой-то из ветвей остается нерешенная молекула (не являющаяся коммерчески доступной), то такое ретросинтетическое планирование считается неуспешным. При этом ретросинтетическое разбиение реализуется справа (от целевой молекулы) налево (к коммерчески доступным соединениям), а получаемая в итоге синтетическая схема – наоборот, слева направо (рисунок 5). Непосредственный синтез целевой молекулы осуществляется согласно синтетической схеме.

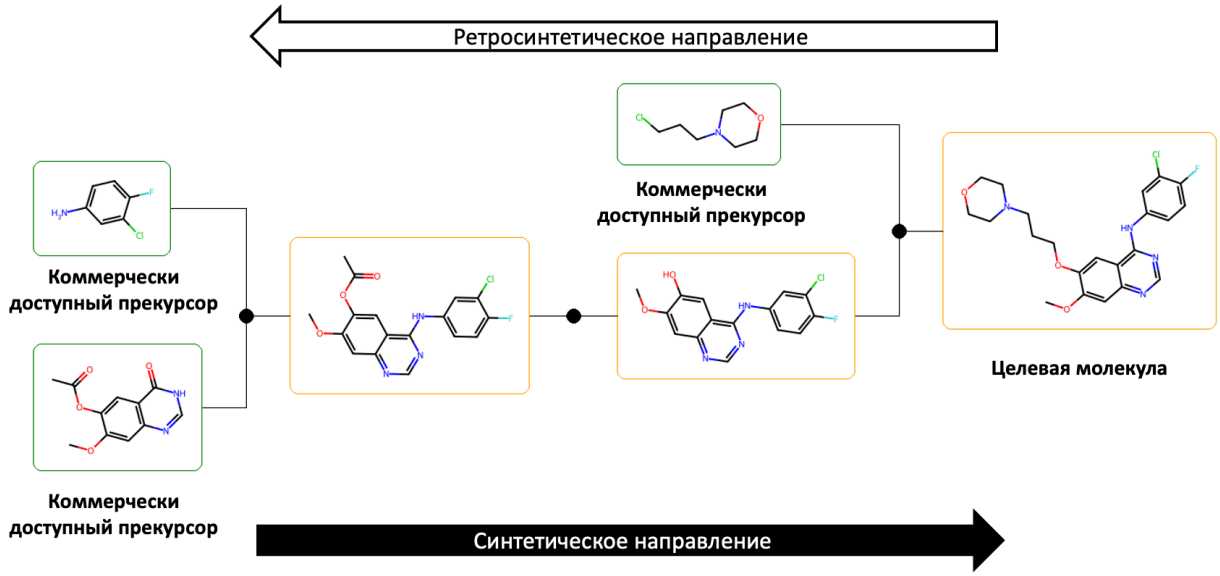

Рисунок 5 – Пример многостадийного синтеза препарата Gefitinib [21]

Для описания качества многостадийного ретросинтетического планирования не существует общепринятых метрик. Качество предлагаемых схем синтеза может быть оценено с помощью двойного слепого AB-тестирования [22]: химику предлагается выбрать одну из схем синтеза молекулы – предсказанную алгоритмом или предложенную другими экспертами / известную в литературе. Кроме того, алгоритмы и предлагаемые ими схемы часто сравнивают с точки зрения скорости предсказания, количества решенных молекул, средней длины схем и других параметров.

1.3 Подходы к ретросинтезу на основе МО и ИИ

Ключевыми задачами ретросинтеза, решаемыми с помощью МО и ИИ, являются предсказание одностадийной ретросинтетической реакции и многостадийное ретросинтетическое планирование (см. подраздел 1.2). Каждая из этих задач включает в себя большое количество подзадач, которые необходимо решить для качественного предсказания одностадийной

ретросинтетической реакции или полной ретросинтетической схемы (рисунок 6).

Рисунок 6 – Задачи и подзадачи ретросинтеза [9]

Очевидно, что наиболее комплексной задачей с точки зрения реализации и при этом наиболее полезным инструментом для химиков синтетиков представляется многостадийное ретросинтетическое планирование, являющееся объектом исследования данной работы. В текущем подразделе наибольшее внимание будет уделено инструментам для предсказания именно многостадийных синтетических и ретросинтетических схем на основе искусственного интеллекта и машинного обучения.

Одними из первопроходцев в использовании искусственного интеллекта и машинного обучения для ретросинтетического анализа стали авторы Сеглер и Уоллер в 2017 году. В своей статье «Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction» [23] они представили первый пример использования символических нейронных сетей для предсказания химических

и ретросинтетических реакций. В рамках данной работы молекулы, представленные в виде молекулярных дескрипторов ECFP4 [24], передавались в предобученную однослойную нейронную сеть FC512 ELU [25]. Нейросеть предсказывала наиболее подходящую именную реакцию (шаблон), с помощью которой данная молекула может быть разбита, а затем наиболее вероятный шаблон применялся к целевой молекуле и предлагал конкретные реагенты (рисунок 7).

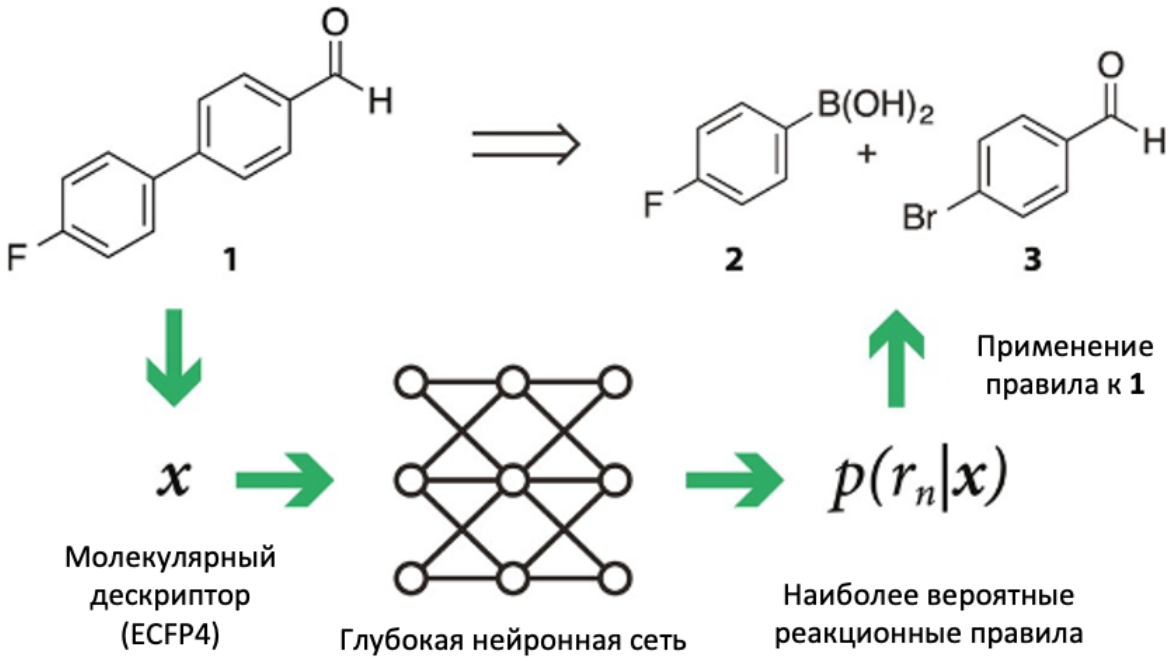

Рисунок 7 – Схема получения ретросинтетического предсказания с помощью нейронной сети [23]

Для обучения и валидации нейросети из базы данных Reaxys [13] было извлечено примерно 4.9 миллиона реакций (известных до 2015 года) с одним продуктом и максимум тремя реагентами. Полученный датасет был разбит на тренировочную, тестовую и валидационную выборки в соотношении 7:1:2. Таким образом, использовав для обучения модели около 3.5 миллионов реакций авторам удалось добиться метрики Top10Acc в 95% для предсказания

ретросинтетической реакции и 97% для предсказания химической реакции на валидационном датасете в почти 1 миллион реакций. Для сравнения, в тех же условиях экспертные системы показывали метрику Top10Acc не более 19% для обоих типов предсказаний.

Ключевым преимуществом нейросетевого подхода авторы отмечают отсутствие необходимости в ручном введении реакционных правил, что особенно важно для сложных реакций с несколькими возможными реакционными центрами. Было показано, что их нейросеть способна самостоятельно учитывать молекулярный контекст и предлагать реакции, селективно идущие только по необходимому в конкретном случае реакционному центру.

Для иллюстрации качества работы модели авторы выбрали отсутствующую в тренировочной выборке молекулу и построили для неё цепочку ретросинтетических разбиений, рекурсивно передавая в модель лучшее предсказание с предыдущей стадии (рисунок 8). Полученное в итоге разбиение соответствовало литературной схеме синтеза данной молекулы [26].

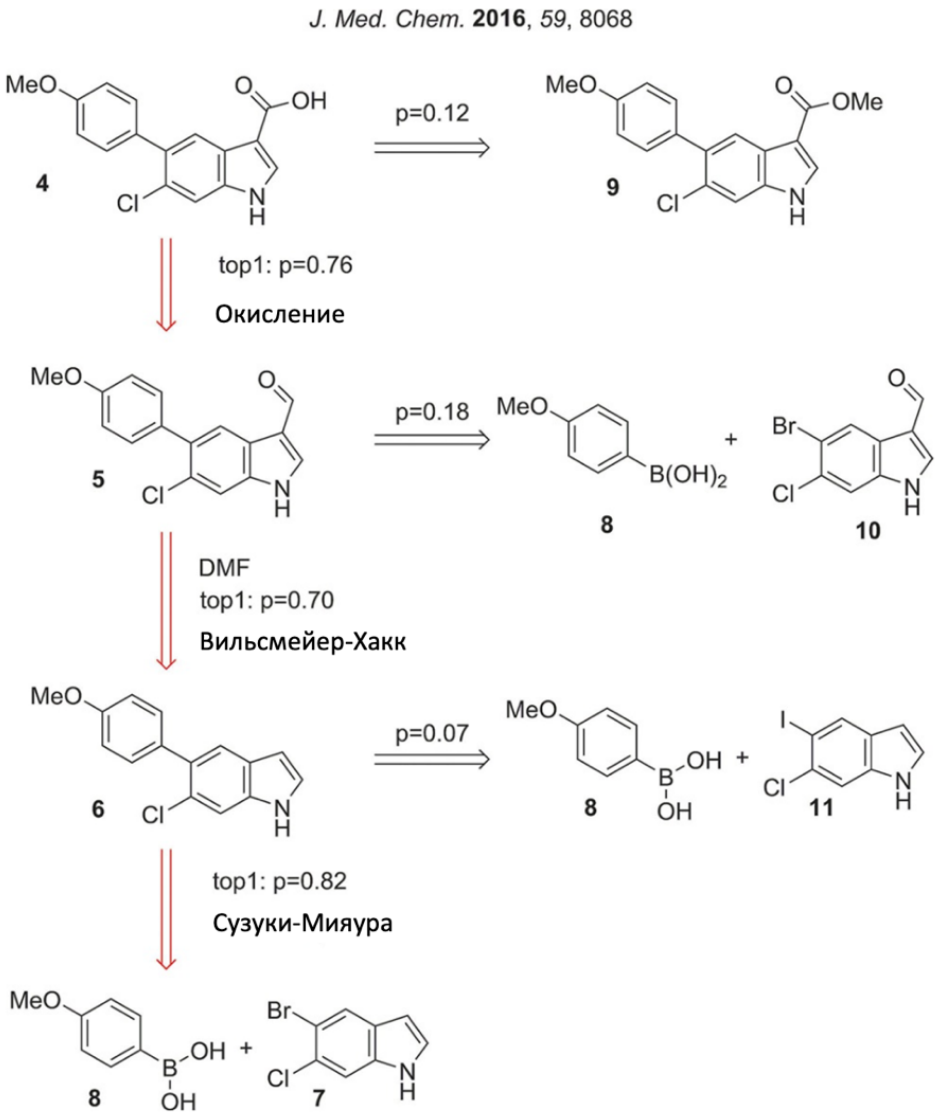

Рисунок 8 – Рекурсивное применение нейросети для ретросинтеза [23], отмеченная красным ветка соответствует литературному пути [26]

Тем не менее, простое рекурсивное использование лучших предсказаний не является оптимальным способом построения многостадийных ретросинтетических схем. Чем сложнее молекула и, соответственно, длиннее потенциальная схема её синтеза, тем больше существует вариантов построения полной схемы. Выбор оптимального пути синтеза среди огромного множества возможных реакционных последовательностей требует значительно более комплексного подхода, который должен учитывать

доступность реагентов, реализуемость конкретной цепочки превращений, постановку/снятие защитных групп и многое другое.

В своей следующей статье «Planning chemical syntheses with deep neural networks and symbolic AI» 2018 года [22] упомянутые выше Сеглер и Уоллер развили свой нейросетевой подход и представили новый способ предсказания многостадийного ретросинтетического анализа с использованием комбинации трех различных нейросетей и алгоритма поиска по дереву Монте-Карло (Monte-Carlo Tree Search, MCTS) [27]. Авторы называют такой подход 3N-MCTS (рисунок 9).

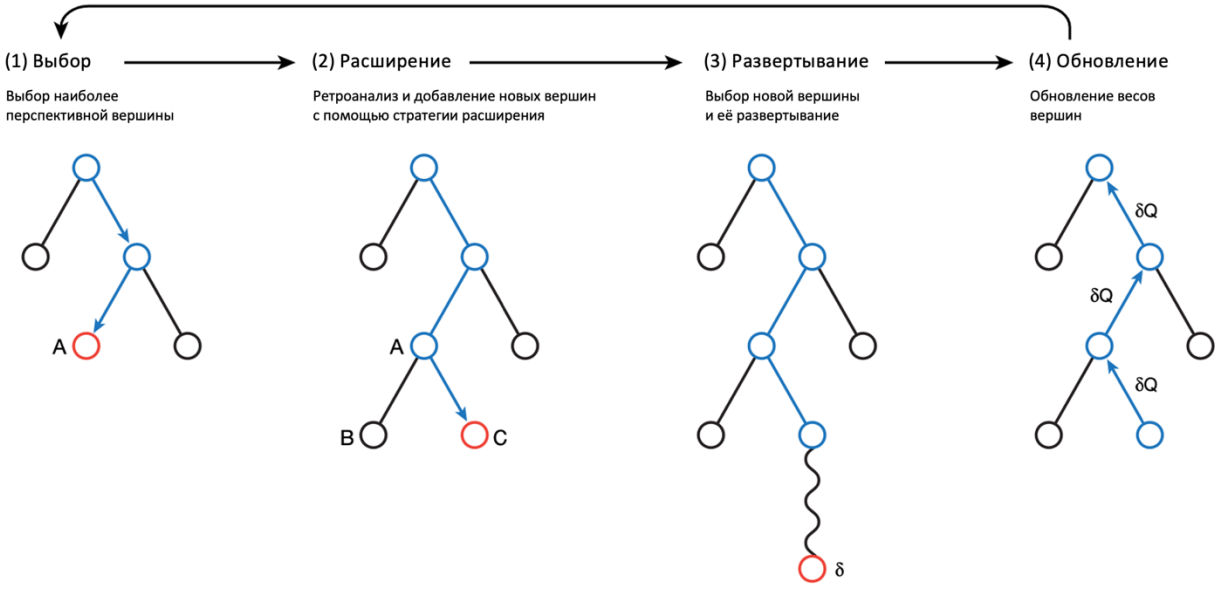

Рисунок 9 – Алгоритм поиска по дереву методом Монте-Карло [22]

Алгоритм Монте-Карло отлично зарекомендовал себя в задачах поиска оптимальной цепочки решений в слишком больших для полного перебора пространствах состояний (например, игра го и алгоритм AlphaGo) [28–30], к которым также можно отнести и пространство всех возможных методов синтеза целевой молекулы. Данный алгоритм состоит из следующих четырех стадий, включающих использование трех нейросетей:

1. Выбор (Selection)

На первом этапе алгоритму необходимо выбрать наиболее перспективную с точки зрения химического превращения вершину дерева. Если такая вершина посещается впервые, то она сразу переводится на стадию развертывания 3 для оценки, в противном случае запускается стадия расширения 2. Таким образом алгоритм производит балансировку между исследованием непроверенных узлов и использованием узлов с высокими оценками.

2. Расширение (Expansion)

На втором этапе в работу включается первая нейросеть (стратегия расширения), предлагающая набор шаблонов реакций для получения узлового соединения. Далее отбирается k наиболее вероятных шаблонов, с помощью которых генерируются возможные реагенты для осуществления реакции. Полученные реакции проверяются второй нейросетью (стратегия фильтрации), оценивающей возможность протекания химического превращения с данными реагентами. В случае успешного предсказания соответствующие реагенты добавляются в качестве новых вершин дерева (рисунок 10).

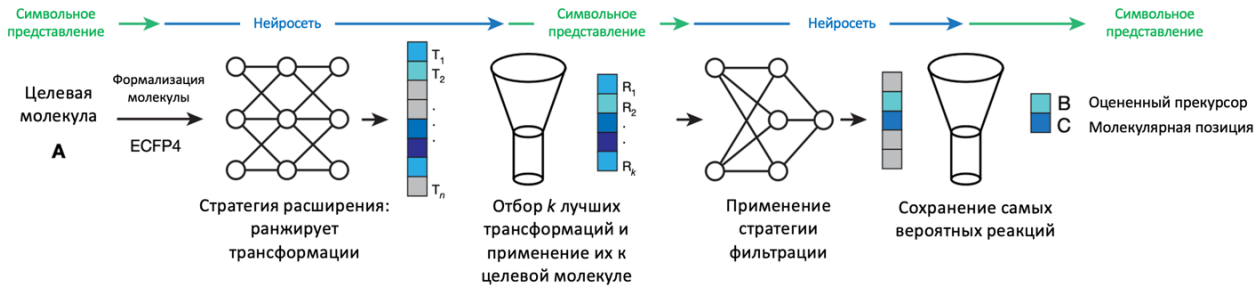

Рисунок 10 – Стадия расширения (2) в алгоритме Монте-Карло [22]

3. Развертывание (Rollout)

На третьем этапе происходит последовательное развертывание вглубь вершин, добавленных на этапе расширения. Здесь вступает в игру третья

нейросеть (стратегия развертывания) со схожей, но значительно более простой архитектурой, чем у первой нейросети. В отличие от стратегии расширения, стратегия развертывания выдает только одно лучшее предсказание и поэтому работает значительно быстрее. Она применяется последовательно до тех пор, пока не достигнет ограничения по длине схемы или коммерчески доступных соединений.

4. Обновление (Update)

На четвертом этапе происходит переоценка узлов. Если в ходе стадии развертывания решение было найдено, то данный узел получает оценку 1; если не было найдено ни одного решения – выставляется оценка минус 1; в случае частичного решения выставляется промежуточная оценка, рассчитываемая по специальной формуле (которая при желании может быть заменена на собственную). В результате осуществляется обратное распространение обновленных оценок для узлов, после чего происходит переход к первому этапу.

Эти четыре фазы повторяются до тех пор, пока не будет достигнуто ограничение по времени или числу итераций. Для построения итогового синтетического плана последовательно выбираются узлы с максимальной оценкой, пока не будет достигнут решенный узел или превышена максимальная глубина схемы. В последнем случае схема будет считаться нерешенной.

Авторы отмечают, что разработанная ими система для многостадийного ретросинтетического планирования в два раза более эффективна с точки зрения количества верных решений и в тридцать раз быстрее классических экспертных систем на основе вручную введенных правил. Качество работы алгоритма оценивалось на примере 9 молекул с помощью двойного слепого АВ-тестирования, описанного в подразделе 1.2. Данное тестирование показало, что в большинстве задач химики чаще отдавали предпочтение синтетическим схемам, предсказанным алгоритмом 3N-MCTS.

Несмотря на все достоинства решения, представленного в статье Сеглера и Уоллера, данная работа оказалась не лишена существенных недостатков. Например, в рамках статьи не приведено описание обработки данных о химических реакциях и обучении нейросетей. Кроме того, авторы представили только лишь концепцию решения без готовой реализации, что делает затруднительным тестирование и использование описанного инструмента в собственных задачах. В дополнение хочется отметить коммерческий характер распространения данных для обучения нейросетей на основе базы Reaxys, делающий эти данные недоступными для свободного использования.

В 2020 году, основываясь на описанной выше статье, группа разработчиков из компании AstraZeneca [31] опубликовала работу «AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning» [32], в которой представила свою реализацию инструмента для многостадийного ретросинтетического планирования на основе использования алгоритма Монте-Карло и нейронных сетей. Стоит отметить, что название статьи отлично описывает ключевые достоинства AiZynthFinder: это быстрый, надежный и гибкий инструмент, который, что крайне важно, имеет открытый исходный код со свободной лицензией MIT [33].

Главной целью авторов было создание общедоступного и удобно поддерживаемого программного продукта, поэтому за основу был взят один из наиболее популярных языков программирования Python 3. Для взаимодействия с химическими сущностями используется библиотека RDKit [34], нейросетевые модели используют пакет TensorFlow [35].

При написании кода авторы руководствовались принципами объектно-ориентированного программирования (ООП) и модульности системы, что позволило значительно упростить ядро AiZynthFinder. В статье приводится сравнение с open-source решением ASKCOS [36]: на момент написания статьи в ядре AiZynthFinder содержалось 1095 инструкций, тогда как в ядре ASKCOS

– 2336 инструкций. Всё это делает AiZynthFinder очень удобным с точки зрения продолжительной поддержки и расширения функциональности: на данный момент было выпущено более 20 релизов, среди которых четыре мажорных версии (последняя вышла в декабре 2023 года) [37]. Кроме того, код имеет обширную документацию и множество интерфейсов (Python, Jupiter Notebook GUI, cli-утилита), позволяющих взаимодействовать с AiZynthFinder пользователям с разными техническими уровнями.

С точки зрения быстродействия и качества предсказаний AiZynthFinder не уступает, а во многих аспектах даже превосходит ASKCOS. При тестировании на 100 молекулах из базы данных ChEMBL [38] было показано, что в среднем инструмент от AstraZeneca был в 2-3 раза быстрее своего конкурента, хотя и показал чуть меньшее количество решенных молекул: 55 против 62 (см. таблицу 1).

Таблица 1 – Сравнение AiZynthFinder и ASKCOS [32]

Параметр	**AiZynthFinder**	**ASKCOS**
Число инструкций в ядре	1095	2336
Общее число инструкций	1495	9987
Средняя сложность кода	2.2	3.4
База данных реакций	USPTO	Reaxys
Стоки	ZINC	Sigma, eMolecules
Среднее время поиска (с)	38.7	151.0
Среднее время нахождения решения (с)	7.1	14.3
Число решенных молекул (из 100)	55	62
Среднее число реакций	2.4	3.3
Среднее число прекурсоров	2.7	3.2

Разница в результатах инструментов связана с тем, что они работают с разными ретросинтетическими моделями и стоками – наборами соединений, доходя до которых алгоритм прекращает свою работу по дальнейшему расширению схемы. В случае AiZynthFinder модель обучена на открытой базе

химических реакций USPTO [39], а в качестве стоков использует открытую базу данных ZINC [40], в то время как ASKCOS обучен на данных Reaxys [13] и использует стоки от Sigma и eMolecules [41]. Авторы указывают, что при использовании одинаковых стоков и моделей, обученных на одних и тех же наборах данных, алгоритмы демонстрируют схожее количество решенных схем. Главным недостатком инструмента ASKCOS авторы отмечают его ориентированность на веб-версию и вычисления на собственных серверах компании-производителя, что и правда делает использование данного продукта несколько сомнительным с точки зрения информационной безопасности.

В дополнение к ретросинтетическому инструменту авторы прикладывают скрипты для создания стоков и обучения ретросинтетических моделей. Всё это делает AiZynthFinder очень удобным и гибким в плане тонкой настройки под нужды конечного пользователя.

Одним из недавних open-source решений для ретросинтетического планирования является работа 2023 года «Retrosynthesis prediction using an end-to-end graph generative architecture for molecular graph editing» [42]. В данном исследовании была разработана модель предсказания ретросинтеза Graph2Edits, которая пошагово предсказывает возможное изменение графа продукта, создавая промежуточные синтоны и исходные вещества. Модель отличается от других подходов, поскольку не использует шаблоны реакций. Вместо этого архитектура Graph2Edits, основанная на графовой нейронной сети (GNN), рассматривает одностадийный ретросинтез как последовательность изменений графа и, по словам авторов, генерирует реагенты в такой же манере, как это делают химики-синтетики.

Метод продемонстрировал метрику Accuracy (Top1Acc) в 55.1% на датасете USPTO-50k, что сопоставимо или лучше, чем у самых современных моделей. На более широком и шумном датасете USPTO-full достигнута

точность в 44.0%, что также близко к лучшим моделям. Результаты указывают на хорошую обобщающую способность и устойчивость метода Graph2Edits.

Для подтверждения работоспособности своей модели авторы выбрали три известных молекулы и попытались воспроизвести литературные схемы их синтеза, последовательно применяя модель на каждой стадии. В результате, в большинстве случаев правильное разбиение соответствовало лучшему предсказанию, хотя и было несколько стадий, для которых соответствие обнаруживалось только для шестого или восьмого по счету предсказания.

Несомненным плюсом данной работы является тот факт, что авторы приложили открытый код модели на GitHub [43]. Тем не менее, представленный код не имеет обширной документации и обновлялся ровно один раз – в момент публикации статьи. Готовым решением это назвать сложно, хотя и приложенные данные представляют исследовательский интерес. Кроме того, инструмент Graph2Edits не включает в себя возможность автоматического построения многостадийной ретросинтетической схемы, что делает менее привлекательной его имплементацию в качестве полноценного продукта для химиков синтетиков.

1.4 Вывод

На сегодняшний день существует огромное множество алгоритмов ретросинтеза на основе машинного обучения и искусственного интеллекта, в рамках литературного обзора была рассмотрена лишь малая их часть. Для примера, в обзорной статье «A Unified View of Deep Learning for Reaction and Retrosynthesis Prediction: Current Status and Future Challenges» [11] подробно сравниваются почти 40 моделей для предсказания одностадийных химических и ретросинтетических реакций. Напротив, согласно обзорной статье [9], инструменты для многостадийного ретросинтетического планирования представлены в значительно меньшем количестве. В своем обзоре авторы

проводят сравнение для всего лишь шести таких продуктов, каждый из которых имеет ряд особенностей (таблица 2).

Таблица 2 – Сравнение инструментов для многостадийного планирования [9]

Инструмент	**Доступность**	**Пользовательский интерфейс**	**Особенности**
AiZynthFinder [31]	Бесплатный	Python / GUI в Jupyter notebook	Код в свободном доступе, хорошая документация
ASKCOS [35]	Бесплатный	Веб-сервис (есть возможность локального запуска)	Датасет Reaxys, высокая ресурсоемкость, сложная архитектура
RoboRXN [44]	Бесплатный	Облачный сервис, API	Запуск только на серверах поставщика, API недоступен в России
Synthia [15]	Коммерческий	Веб-страница, ввод рисованием молекул	100 000 вручную прописанных правил
SciFinder [14]	Коммерческий	Веб-страница, ввод рисованием молекул	Интеграция с собственной базой данных химических реакций
Spaya AI [43]	Коммерческий	Веб-страница, ввод SMILES, API	Возможность анализа тысяч молекул с помощью API

В рамках данной выпускной квалификационной работы ключевыми требованиями к ретросинтетическому инструменту являются его доступность, возможность локального запуска и тонкой настройки под свои нужды. Среди представленных шести фреймворков половина распространяется на коммерческой основе и, к сожалению, недоступна в России (Synthia [16], SciFinder [15], Spaya AI [44]). Из оставшихся трех бесплатных инструментов два (ASKCOS [36], RoboRXN [45]) в основном рассчитаны на работу с помощью проприетарных веб-сервисов, что делает запуск предсказаний для

незапатентованных молекул не очень безопасным. При этом для использования из России на данный момент доступен только ASKCOS. Хотя ASKCOS и имеет возможность локального запуска, данный продукт является достаточно громоздким и ресурсоемким, что значительно усложняет его использование, доработку и внедрение в собственный веб-сервис.

В результате, наиболее привлекательным для имплементации видится инструмент AiZynthFinder [32], поскольку он имеет удобный python-интерфейс, а также хорошо задокументирован и прост в использовании.

Тем не менее, несмотря на все преимущества AiZynthFinder, он имеет несколько аспектов, требующих доработки. Главным фактором, который может повлиять на качество предсказаний, авторы указывают использование более репрезентативных данных для стоков. Кроме того, более тщательный подбор параметров анализа может также положительно повлиять на работу алгоритма.

Всё это делает актуальной цель данной работы, заключающуюся в оценке влияния качества данных и выбора настроек модели машинного обучения на результаты ретросинтетического анализа. В качестве ключевого ретросинтетического инструмента для осуществления данной цели был выбран AiZynthFinder. В следующем разделе будет более подробно рассмотрена техническая сторона реализации данного программного обеспечения.

2 Методическая часть

2.1 AiZynthFinder

В рамках данной дипломной работы был использован пакет AiZynthFinder версии 3.6.0, написанный на языке программирования Python версии 3.9. Ниже перечислены ключевые программные пакеты, на которых основана работа AiZynthFinder:

– Модули RDKit [34] / RDChiral [46] ответственны за формализацию химической части. С их помощью производятся преобразование SMILES в объект молекулы, работа с шаблонами реакций, генерация InChiKey и многое другое;

– Модули TensorFlow [35] / Keras [47] / Onnx [48] ответственны за работу с нейросетевыми моделями. С их помощью производится создание, обучение и сохранение нейросетей для различных этапов поиска по дереву.

2.1.1 Архитектура

Работа AiZynthFinder состоит из двух ключевых этапов:

– Поиск по дереву, во время которого происходит построение множества схем синтеза согласно алгоритму Монте-Карло;

– Анализ и постобработка, отвечающие за выбор наиболее удачных схем синтеза.

На рисунках 11 и 12 представлены диаграммы взаимоотношений между основными классами AiZynthFinder, которые ответственны за поиск по дереву и анализ / постобработку соответственно. На обоих рисунках, согласно UML-нотации, линия со сплошным ромбом указывает на отношение «владеет», линия с пустым ромбом указывает на отношение «имеет», а пунктирная линия со стрелкой указывает на отношение «использует».

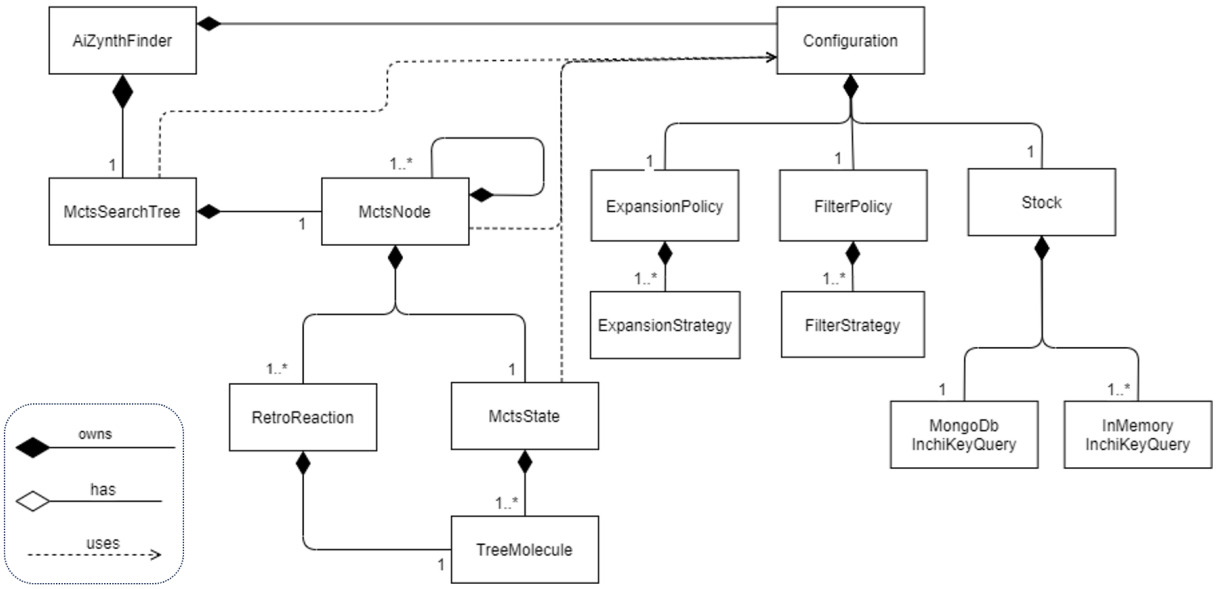

Рисунок 11 – Связи модулей для поиска по дереву в AiZynthFinder [49]

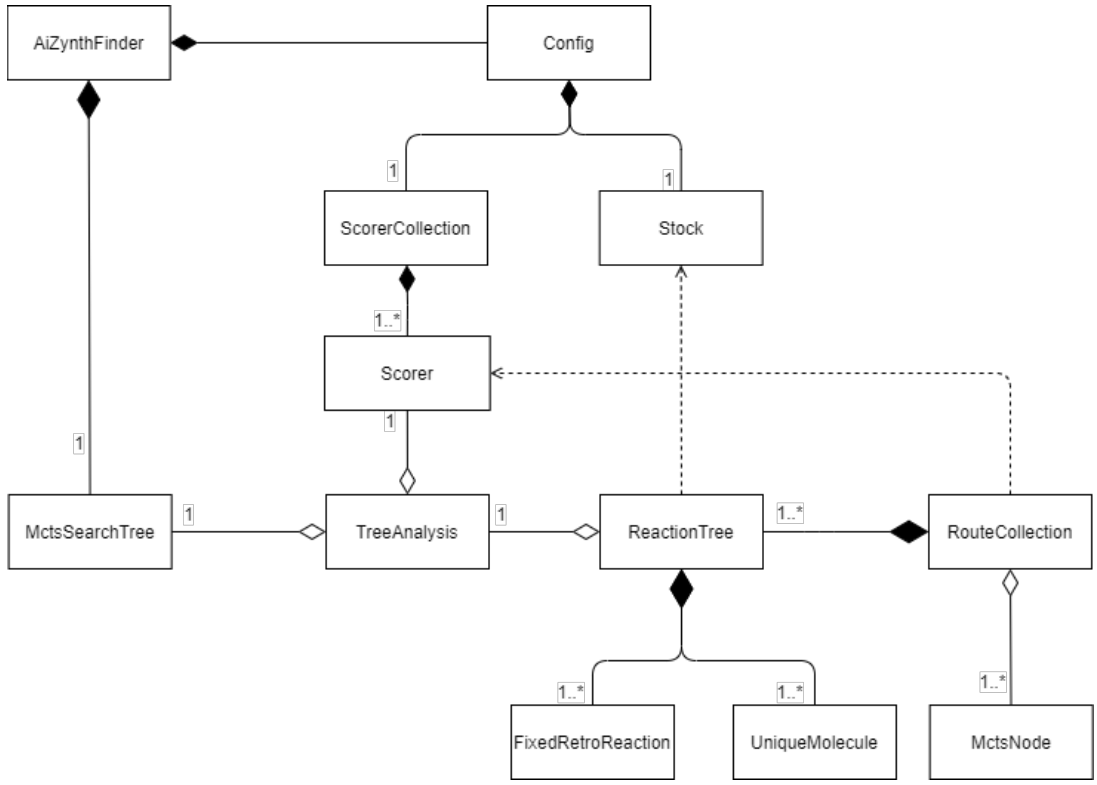

Рисунок 12 – Связи модулей анализа и постобработки в AiZynthFinder [50]

Класс AiZynthFinder является главным объектом, отвечающим за взаимодействие между всеми компонентами анализа. Он загружает

пользовательскую конфигурацию из файла config.yml, в котором хранятся пути к файлам стоков и моделям стратегий разбиения/фильтрации. Полученный объект конфигурации используется для управления поиском по дереву, который осуществляется согласно алгоритму Монте-Карло классом MctsTreeSearch. По окончании поиска по дереву класс AiZynthFinder осуществляет обработку результатов: создается объект RouteCollection, хранящий список сериализованных схем синтеза (ReactionTree), отсортированных согласно использованной скоринговой функции Scorer. Более подробное описание алгоритмической части работы AiZynthFinder можно прочесть в документации [51].

Стоки – это коллекции соединений, доходя до которых алгоритм прекращает дальнейшую декомпозицию целевой молекулы в данной ветке. По сути, соединения из стоков мы считаем доступными прекурсорами для использования в рамках ретросинтетического анализа целевой молекулы. Наиболее предпочтительным выбором для создания стоков являются молекулы из собственного каталога, а также из каталогов поставщиков химической продукции и открытых баз данных.

Стоки могут храниться в виде файлов в hdf5 формате или загружаться из базы данных MongoDB. В любом случае при работе инструмента стоки загружаются в память в виде соответствующего объекта InchiKeyQuery. В сущности, сток представляет собой таблицу, в которой обязательно должен быть столбец со строками хешей InChiKey, соответствующих структурам доступных молекул. В данную таблицу можно добавлять и другие поля, например цену за 1 грамм. В базовой версии AiZynthFinder предоставляет один сток zinc (сокращение – zn) [40], состоящий из 17.42 миллионов молекул, которые представляют собой коммерчески доступные соединения для виртуального скрининга. Авторы отмечают, что для создания данного стока они использовали молекулы с молекулярной массой до 250 г/моль,

липофильностью (logP) до 2.5 и средней или высокой реакционной способностью.

Модели представляют собой файлы формата hdf5 или onnx, хранящие веса соответствующей нейронной сети. В рамках AiZynthFinder существует два вида таких нейросетей: стратегия расширения (более сложная) и стратегия фильтрации (более простая). С пакетом AiZynthFinder поставляется две очень разных стратегии расширения: uspto и ringbreaker. Рассмотрим их подробнее:

1) Uspto (u) – это стандартная стратегия расширения, которая нацелена на последовательную декомпозицию целевой молекулы на так называемые билдинг-блоки. Как правило, в рамках данной стратегии строятся довольно длинные схемы синтеза, наиболее похожие на реальные литературные примеры. Uspto-стратегия сильно зависит от стоков, так как останавливается именно на них, тратя на поиск всё предоставленное время. Основана на датасете, состоящем из 2.33 миллионов реакций и 42.6 тысяч реакционных шаблонов в большинстве своем с линейной топологией;

2) Ringbreaker (rb) – это стратегия расширения, направленная на разбиение циклов. Здесь обычно строятся достаточно короткие схемы (1–3 стадии, редко больше), каждая стадия при этом направлена на разбиение того или иного цикла в рамках имеющейся молекулы. Схемы, как правило, строятся быстро, однако они подходят скорее для «вдохновения» химика-синтетика на то или иное разбиение цикла, чем для непосредственной реализации. От стоков ringbreaker-стратегия зависит в меньшей степени, поиск обычно заканчивается раньше достижения стоковых соединений во всех ветках. Основана на датасете, состоящем из 164.9 тысяч реакций и 5.3 тысяч реакционных шаблонов с циклической топологией.

Стратегия фильтрации так же, как и стратегии расширения, основана на датасете USPTO. В стандартном варианте она используется в комбинации как со стратегией расширения uspto, так и с ringbreaker. Стоит отметить, что упомянутые выше датасеты с реакционными шаблонами также входят в

стандартный пакет AiZynthFinder и необходимы для хранения метаинформации о реакциях. Предоставляются они в виде файлов формата csv.gz и скачиваются вместе с файлами моделей и стоков.

2.1.2 Ввод и вывод инструмента

На вход инструменту пользователь подает целевую молекулу, описанную в виде SMILES. Затем запускается анализ с определенными параметрами ретросинтеза, которые будут более подробно расписаны в пункте 2.1.3. По окончании анализа можно получить несколько видов итоговых результатов:

1) RouteCollection – это список схем синтеза, расположенных в порядке убывания оценки схемы синтеза (state score – метрика от ноля до единицы, рассчитываемая алгоритмом и учитывающая количество стадий и прекурсоров в стоках). Сами схемы представляют собой словари с древовидной структурой, состоящие из словарей для молекул и реакций, которые последовательно вкладываются друг в друга с помощью поля "children". При этом наследником молекулы может быть только реакция, а для реакции – только молекула или список молекул. Поле "in_stock" показывает, находится ли данная молекула в стоке или нет. На рисунке 13-a представлен упрощенный вид такого словаря для одностадийной схемы синтеза (для краткости и наглядности некоторые поля были убраны, а ключи – изменены). На рисунке 13-b тот же словарь представлен в виде схемы. Молекула считается успешно решенной, если есть хотя бы одна схема, для которой во всех ветках расширение дошло до молекул из стоков;

2) Изображение схем синтеза – это человекочитаемые картинки схем синтеза, сохраняемые в формате png (рисунок 13-c). В цветных рамках отображаются молекулярные формулы молекул, причем для молекул в стоке задается зеленый цвет рамки, а для молекул не в стоке – желтый. Черной

точкой обозначается реакция, с помощью которой можно превратить реагенты (слева) в продукт (справа);

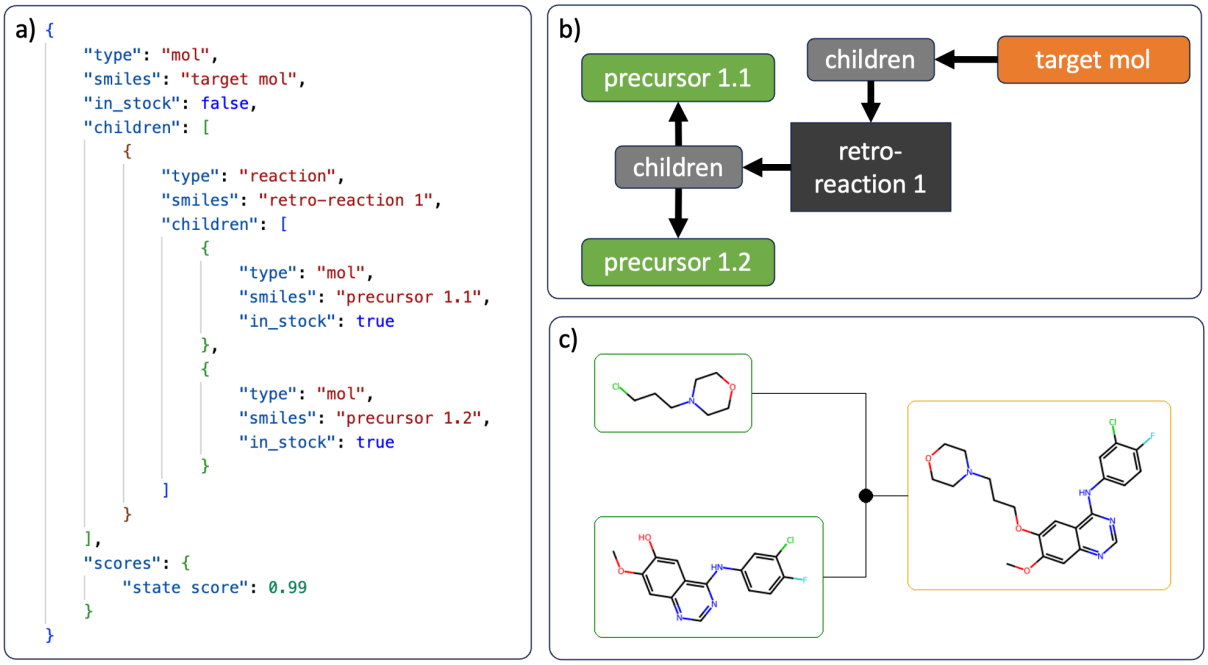

Рисунок 13 – Представление схем синтеза в AiZynthFinder: a) словарь схемы синтеза; b) схема словаря; c) изображение схемы синтеза

3) Статистика анализа – это словарь, хранящий информацию о различных аспектах проведенного ретросинтетического анализа, таких как статус анализа (решена ли молекула), время / итерация нахождения первого решения, число полученных / решенных схем и многое другое (рисунок 14-a). Информация из данного блокнота может быть полезна для оценки качества работы алгоритма;

4) Информация о стоках – это словарь, хранящий информацию о наличии в стоках молекул-прекурсоров в рамках всех полученных схем синтеза (рисунок 14-b). Может быть удобным инструментом для беглого анализа схем синтеза на предмет доступности реагентов (рисунок 14-c).

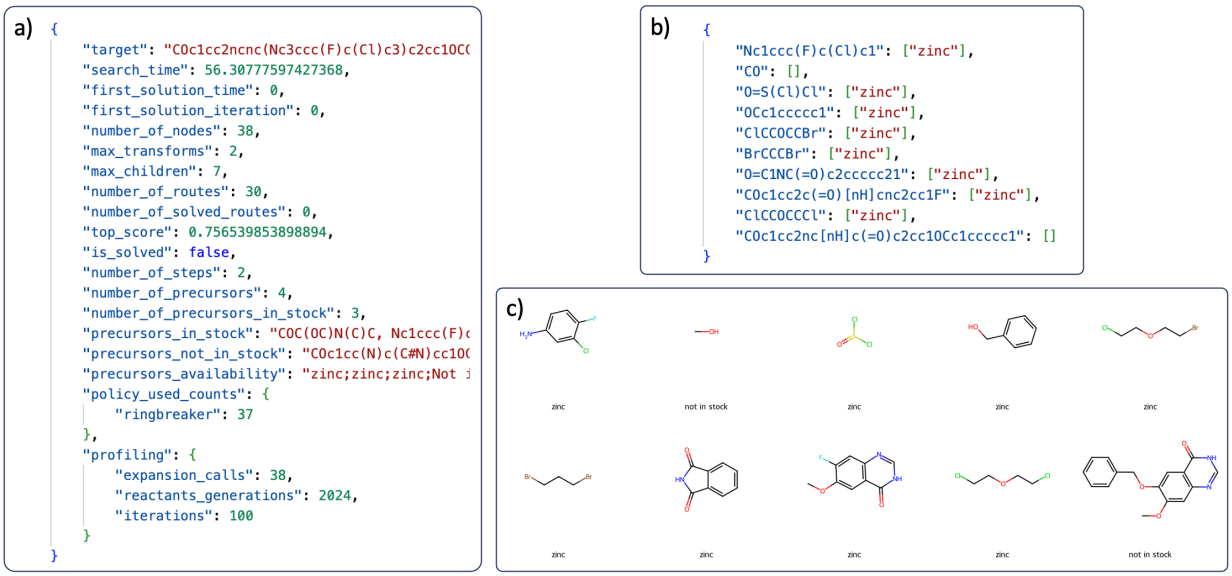

Рисунок 14 – Статистика и информация о стоках: a) словарь со статистикой анализа; b) словарь с информацией о стоках; c) изображение молекул из информации о стоках

2.1.3 Ключевые параметры запуска

AiZynthFinder имеет множество тонких настроек и параметров запуска, которые задаются перед началом анализа и значительно влияют на итоговые результаты. В качестве ключевых параметров можно выделить следующие:

– Target SMILES – строковое представление целевой молекулы, для которой производится ретросинтетический анализ;

– Expansion policy – стратегии расширения или их комбинации. В зависимости от стратегии расширения (uspto или ringbreaker) результаты анализа могут значительно различаться;

– Stock – сток или их комбинации, коллекции доступных соединений, используемых в рамках данного анализа;

– Max transforms (depth) – глубина схемы, максимальное число стадий (химических реакций) в самой длинной ветви схемы синтеза;

– Time – максимальное время, по достижении которого дальнейший анализ прекращается.

Также стоит отметить такие важные параметры, как Iteration (максимальное число итераций поиска по дереву) и Filter policy (стратегии фильтрации).

2.2 Методология оценки качества ретросинтеза

Для оценки качества работы инструмента был создан тестовый датасет из 100 молекул базы данных ChEMBL с определенными характеристиками (рисунок 15). Этот источник был выбран потому, что все молекулы из базы ChEMBL имеют описанный метод синтеза, то есть гипотетически могут быть решены и с помощью инструмента AiZynthFinder. Все использованные в тестовом датасете молекулы являются малыми (с молекулярной массой от 200 до 600 г/моль), находятся на первой фазе клинических испытаний, их липофильность (logP) ограничена в пределах от 2 до 5 (рисунок 16). Применение данных фильтров позволило получить выборку молекул, хорошо соответствующую соединениям из реальной практики работы химического департамента компании Biocad.

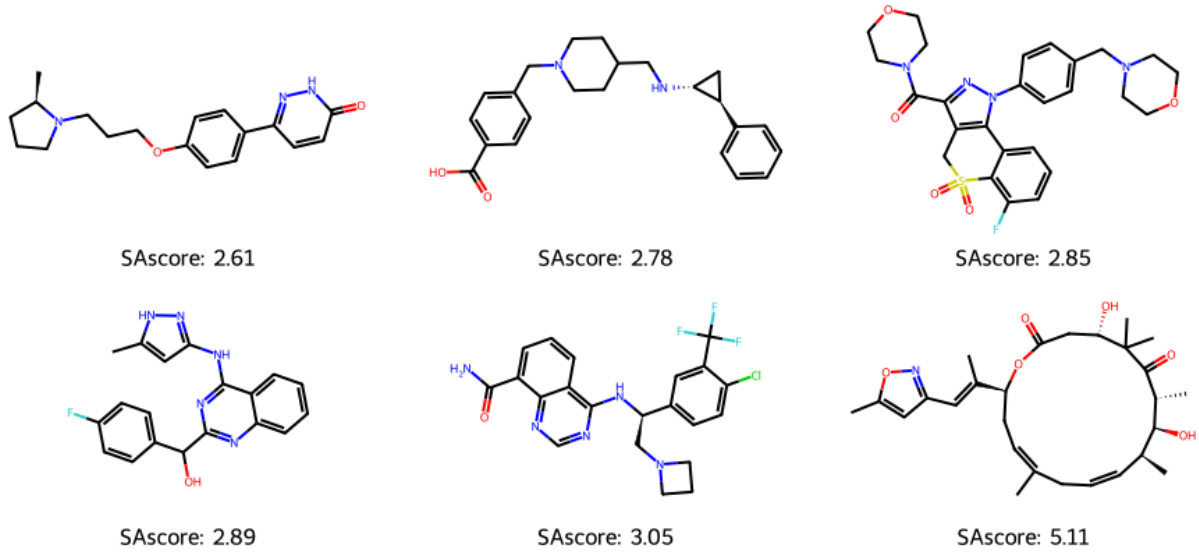

Рисунок 15 – Примеры молекул из тестового набора

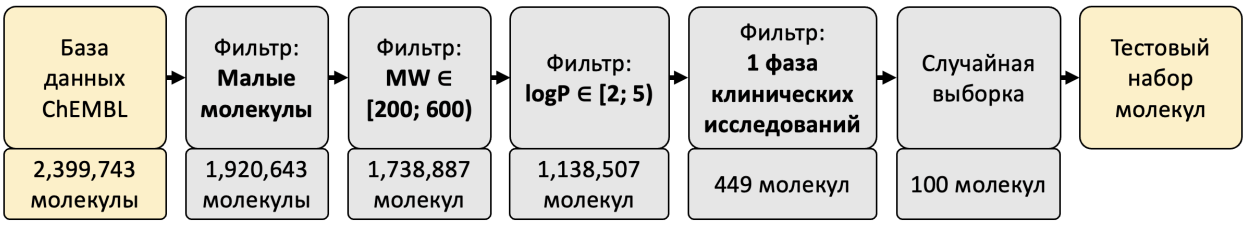

Рисунок 16 – Формирование тестового набора молекул

Кроме того, для дополнительной проверки запуски проводились на датасете AiZynthFinder (рисунок 17), который авторы оригинальной статьи использовали для сравнения работы их инструмента с инструментом ASCKOS. Этот датасет также был собран из известных молекул базы данных ChEMBL, но в данной выборке молекул авторы избавились от стереоцентров – атомов, имеющих как минимум троих различных соседей.

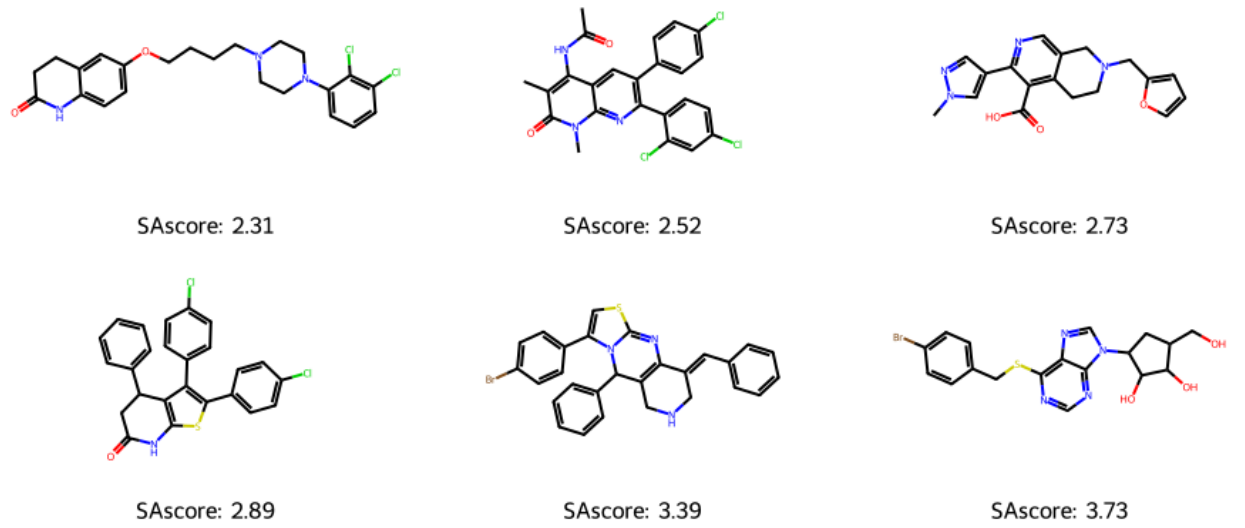

Рисунок 17 – Примеры молекул из набора AiZynthFinder

Для сравнения датасетов по сложности для каждой молекулы была рассчитана метрика SAscore – Synthetic Accessibility, вычисляемая с помощью определенного алгоритма характеристика, описывающая синтетическую доступность молекулы (чем больше SAscore, тем сложнее молекула) [52]. В результате тестовый датасет получился несколько более сложным, чем датасет

AiZynthFinder: средний SAscore составил 3.07 против 2.81 для AiZynthFinder (тест Манна-Уитни, p-value = 0.001).

На рисунке 18 представлена последовательность действий, описывающая процесс проведения экспериментов для оценки качества предсказаний. Полученные наборы по 100 молекул запускали в ретросинтетический анализ на вычислительном кластере с определенными условиями (здесь и далее *условия* – это совокупность стартовых параметров ретросинтетического анализа). По результатам анализа проводился сбор статистики, оценка качества анализа осуществлялась на основании следующих параметров:

– Доля решенных молекул – главная метрика;
– Среднее число стадий;
– Средние значения state score;
– Значения SAscore (для решенных и нерешенных молекул).

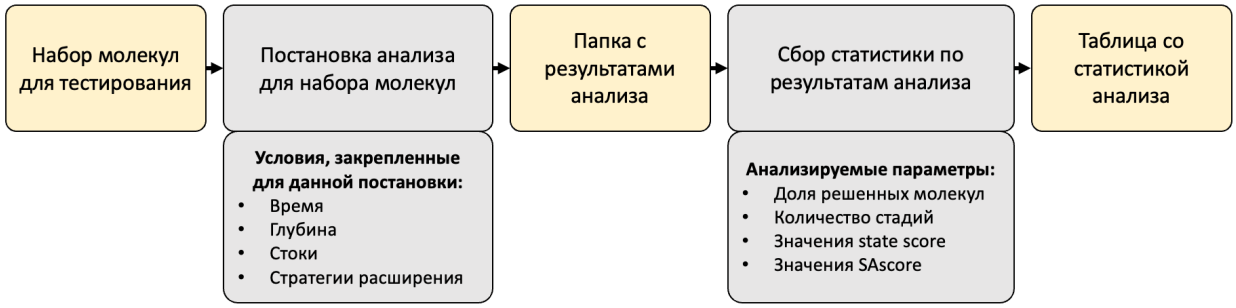

Рисунок 18 – Процесс оценки качества предсказаний

Для запуска множественных предсказаний на вычислительном кластере в различных условиях, а также сбора статистики были написаны специальные скрипты. В качестве варьируемых параметров анализа были выбраны стратегии разбиения (uspto, uspto+ringbreaker), комбинации стоков (см. подраздел 3.3), время анализа (2, 5, 10 и 20 минут) и глубина схемы (6, 9, 12 и 24 стадии). Большая часть экспериментов задействует только стратегию расширения uspto, которая строит наиболее приближенные к реальным схемы

синтеза, a ringbreaker используется только в комбинации с uspto. Стратегия фильтрации и количество итераций были заданы заранее (uspto и 1000 итераций соответственно) и не изменялись в ходе экспериментов.

Ключевые бизнес-требования к инструменту: решение не менее 75% молекул и применимость схем в реальной работе химиков синтетиков. Главной целью оценки качества предсказаний является нахождение оптимальных параметров запуска ретросинтетического анализа, которые можно выставить в качестве стандартных в вычислительном веб-сервисе. Это необходимо для максимального упрощения работы с инструментом и, как следствие, для улучшения пользовательского опыта.

2.3 Вывод

В данном разделе была описана архитектура AiZynthFinder, его ввод/вывод и ключевые параметры запуска. Также была представлена методология оценки качества предсказаний инструмента, ключевой метрикой для которой выбрана доля решаемых инструментом молекул. Главная цель данного исследования – поиск оптимальных условий запуска анализа.

Задача оценки качества работы инструмента будет решаться путем тестирования двух наборов молекул в различных условиях с последующим сбором статистики по результатам анализа. Тестовые наборы представляют собой две непересекающиеся выборки по 100 соединений из базы данных ChEMBL. В каждой выборке содержатся молекулы различной химической природы, что позволяет провести оценку качества работы инструмента как на простых, так и сложных примерах.

3 Обсуждение результатов

3.1 Запуск анализа в стандартных условиях

В AiZynthFinder существует ряд условий запуска анализа, выставляемых в инструменте «по-умолчанию»: стратегия uspto, сток zinc, время 2 минуты и глубина 6. Данные условия были использованы в качестве бэйзлайна. В результате для тестового набора успешно решилась 31 молекула, а для набора AiZynthFinder – 40. Интересно, что в случае набора AiZynthFinder результаты заметно отличаются от данных, предоставляемых в статье [32]: в ней авторы указывают на решение 55 молекул со стоками zinc.

Рассмотрим некоторые наиболее распространенные ошибки алгоритма, приводящие к отсутствию решения для молекулы:

– Недостижение максимальной глубины схемы: скорее всего алгоритм упирается в ограничение по времени анализа, бросая дальнейшее расширение схемы «на полпути» (рисунок 19);

– Недостроенные схемы: алгоритм доходит до ограничения по глубине, но самый продвинутый прекурсор не находится в стоках и, гипотетически, может быть разбит на более простые компоненты в несколько дополнительных стадий (рисунок 20);

– Отсутствие некоторых простых прекурсоров: зачастую алгоритм строит в целом хорошую схему синтеза, однако ему не хватает одного или нескольких достаточно простых и распространенных реагентов для успешного завершения схемы. К таким реагентам могут относиться галогенирующие агенты, простые амины, кислоты и другие соединения, имеющиеся в большинстве лабораторий органического синтеза на постоянной основе (рисунок 21).

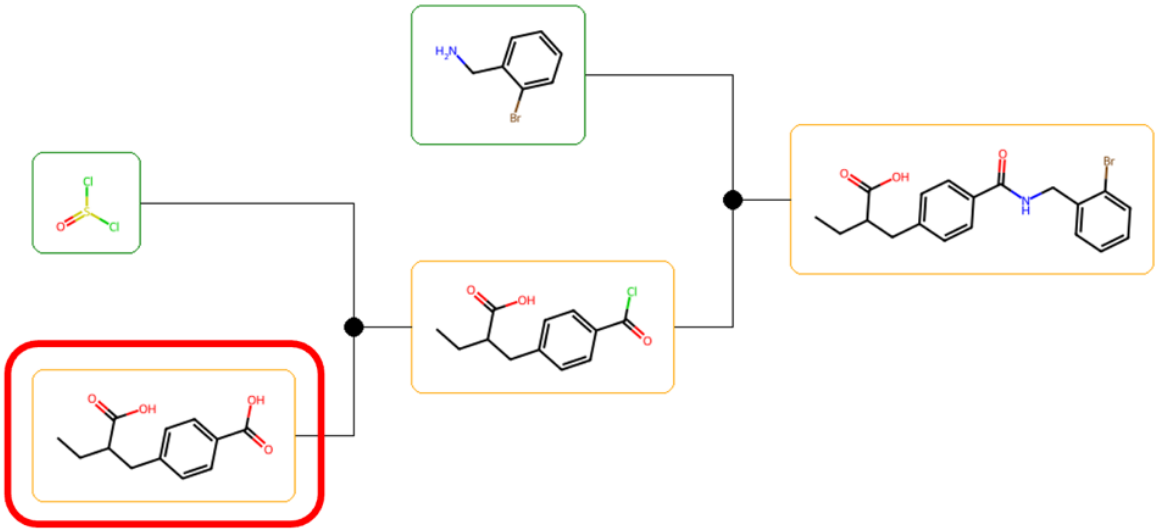

Рисунок 19 – Алгоритм «уперся» в ограничение по времени

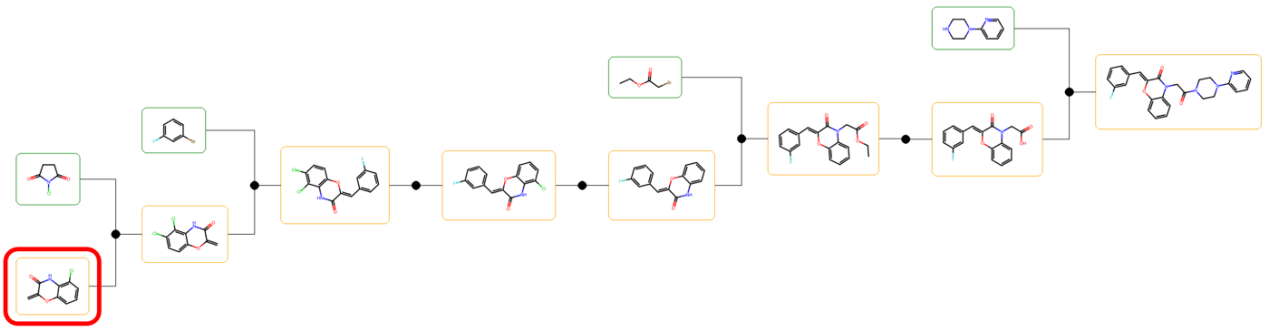

Рисунок 20 – Алгоритм «уперся» в ограничение по глубине

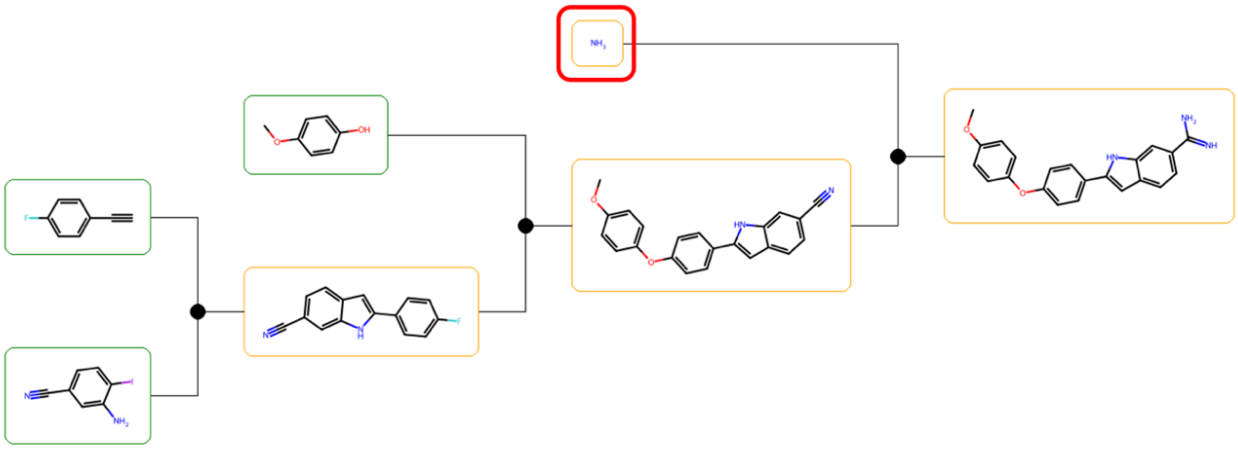

Рисунок 21 – Алгоритм не нашел в стоках простейший реагент

На основании рассмотренных выше ошибок становятся ясны основные направления для улучшения работы алгоритма:

– Варьирование времени анализа;
– Варьирование глубины схем;
– Добавление новых стоков.

3.2 Влияние простых параметров запуска

В данном подразделе будет рассмотрено влияние таких простых параметров анализа, как время и глубина. Для этого проведено 32 запуска анализа для двух наборов по 100 молекул (16 запусков на каждый набор), в рамках которых варьировались время (2, 5, 10, 20 минут) и глубина (6, 9, 12, 24 стадии). Во всех запусках в качестве стока использовался zinc, а в качестве стратегии разбиения – uspto.

3.2.1 Время анализа

Время анализа – это ключевой простой параметр запуска ретросинтеза. Довольно логичной кажется гипотеза, что при увеличении количества времени для анализа будет увеличиваться и число решенных молекул. Связано такое предположение в первую очередь с характером работы алгоритма поиска по дереву Монте-Карло: чем больше мы ему даем времени на построение схем, тем большее количество итераций он сможет произвести, находя всё более выигрышные пути синтеза.

На рисунке 22 представлены графики зависимости доли решенных молекул от времени анализа.

В случае тестового набора обнаруживается следующая зависимость: при увеличении времени анализа растет и число решенных молекул, но чем больше времени мы даем, тем меньший прирост получаем. Так, для 2 минут в

среднем получаем 29% решенных молекул, для 5 минут – 34%, для 10 минут – 38.5% и, наконец, для 20 минут – 43.25%. Такая тенденция наблюдается для всех значений глубины поиска.

Для набора AiZynthFinder наблюдается похожая картина роста числа среднего числа решенных молекул с увеличением времени, но не во всем временном диапазоне, а с 2 до 10 минут: 2 минуты – 38.25%, 5 минут – 40.75% 10 минут – 45.25%. При переходе же к 20 минутам средняя доля решенных молекул не растет, а даже незначительно падает до 45%.

Из сравнения двух наборов молекул между собой можно сделать вывод, что набор AiZynthFinder решается лучше, чем тестовый. Особенно эта разница заметна в случае 2 минут на анализ, но с увеличением времени анализа она нивелируется.

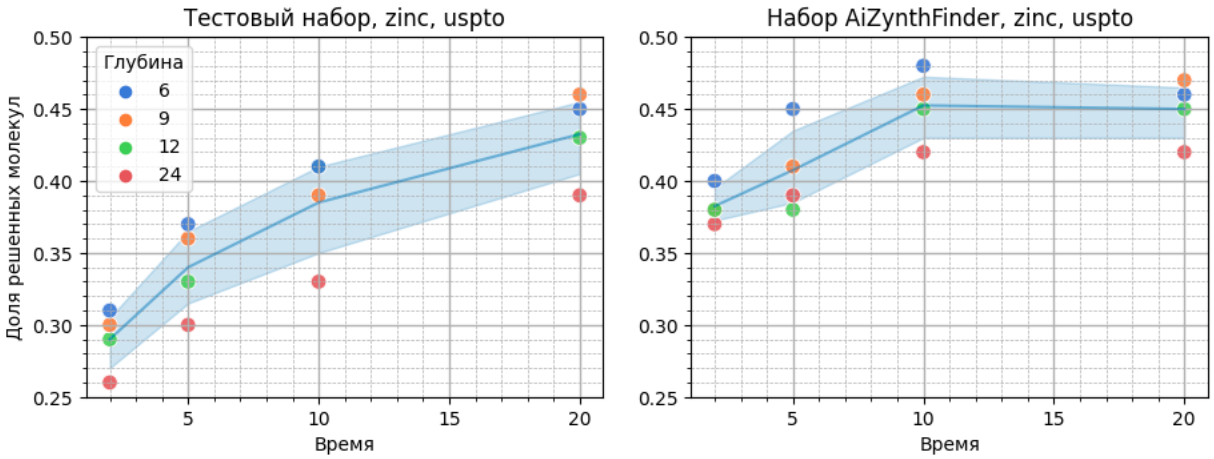

Рисунок 22 – Зависимость доли решенных молекул от времени анализа

3.2.2 Глубина анализа

Глубина также имеет важное значение для результатов ретросинтетического анализа. В зависимости от глубины схемы должен значительно меняться её размер и общий вид. Кроме того, в некоторых случаях схемы не решались из-за достижения максимальной глубины поиска, хотя

нерешенные прекурсоры еще могли быть разбиты на более простые доступные предшественники. Поэтому была выдвинута гипотеза, что при увеличении глубины схемы могут быть получены более высокие доли решенных молекул.

Тем не менее, по результатам экспериментов была обнаружена противоположная тенденция: с увеличением глубины схемы для обоих наборов наблюдалось некоторое понижение среднего числа решенных молекул, снова более выраженное для тестового набора (рисунок 23).

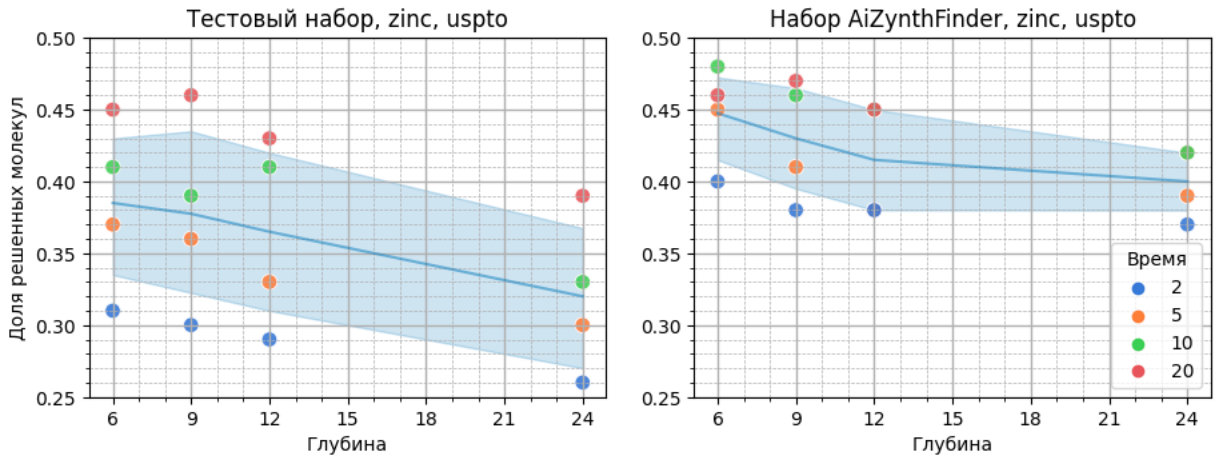

Рисунок 23 – Зависимость доли решенных молекул от глубины

Если же обратиться к зависимости среднего числа реакций на схему от глубины (рисунок 24), то обнаруживается прямая и почти линейная взаимосвязь данных величин. В случае времени значительного влияния на среднее число реакций обнаружено не было.

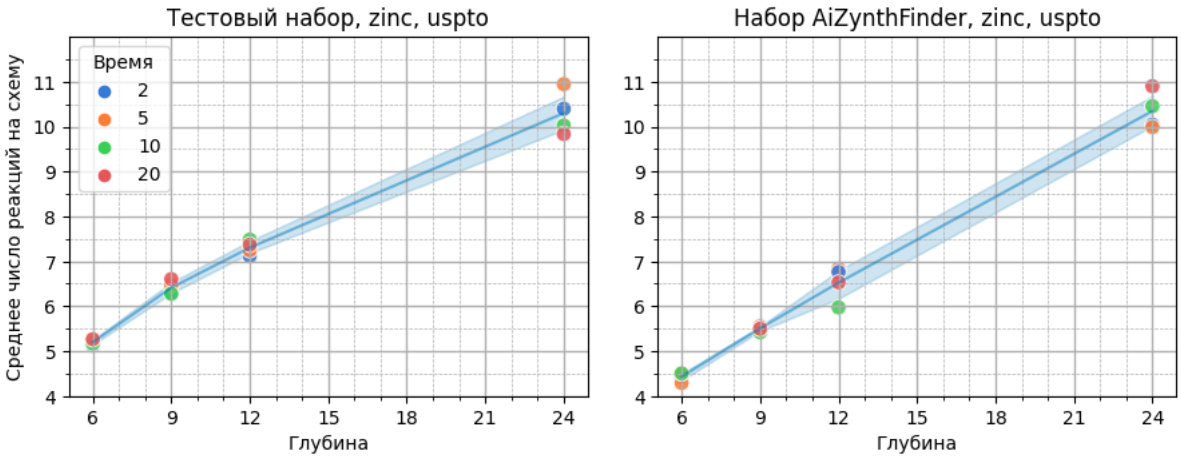

Рисунок 24 – Зависимость доли решенных молекул от глубины

При более внимательном просмотре схем было обнаружено, что увеличение глубины поиска часто приводит к более вероятному проявлению «галлюцинаций» алгоритма: ситуаций, когда вместо разумного разбиения молекулы на более простые компоненты происходят неуместные точечные изменения, фактически не влияющие на потенциал её дальнейшей декомпозиции. На рисунке 25 представлен пример такого поведения (время 2 минуты, глубина 24): вместо разделения молекулы по амидной связи, которое бы привело к значимому упрощению компонентов, происходит бесполезная перезащита гидроксильных групп без значимого изменения кора.

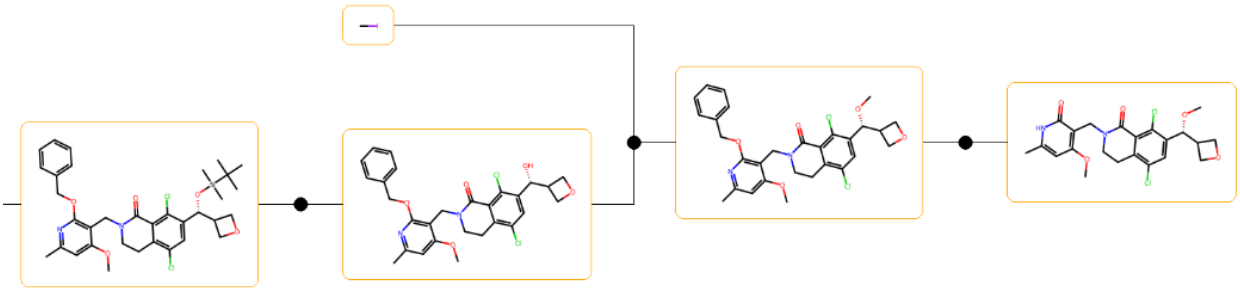

Рисунок 25 – Пример «галлюцинации» алгоритма при построении схемы

Скорее всего, подобный характер построения схем связан с тем, что в отсутствии строгого ограничения глубины при поиске по дереву приоритет

отдается не перебору различных путей, а удлинению зачастую не самых выигрышных веток. Простым молекулам, решаемым за 2–3 стадии, увеличение глубины обычно не мешает, но для более сложных молекул это может стать критичным фактором, приводящим к отсутствию решения.

3.2.3 Выводы по простым параметрам

Исходя из рассмотренных выше данных, можно сделать вывод, что и время, и глубина анализа являются важными параметрами, в значительной степени влияющими на итоговые результаты и вид получаемых схем синтеза (рисунок 26).

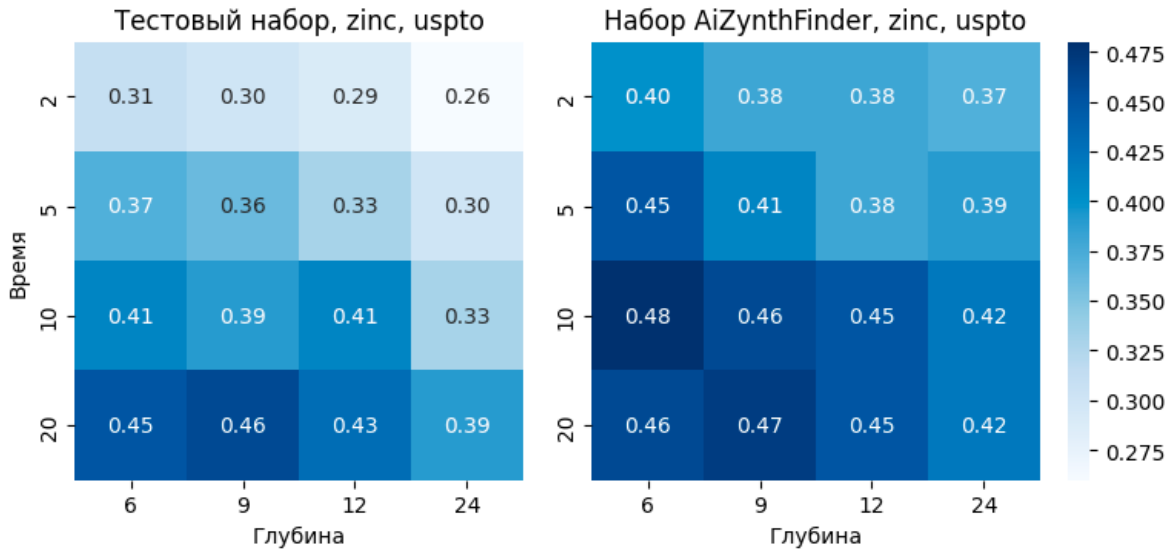

Рисунок 26 – Зависимость доли решенных молекул от времени и глубины

С точки зрения времени общая тенденция сводится к тому, что с увеличением данного параметра растет и доля решенных молекул. Однако, на примере набора AiZynthFinder мы видим, что переход от 10 к 20 минутам практически не повлиял на долю решенных молекул, а в случае тестового набора прирост доли решенных молекул оказался несоразмерен приросту времени. Кроме того, с точки зрения конечного пользователя слишком долгое

ожидание результатов анализа негативно сказывается на опыте использования инструмента, особенно при условии получения одних и тех же результатов для многих молекул. Поэтому оптимальным временем анализа представляются 10 минут.

С точки зрения глубины поиска тенденция также довольно понятна: слишком сильное увеличение данного параметра негативно сказывается на качестве результатов анализа. Разумным выглядит и предположение о том, что снижение глубины до минимального уровня также не приведет к улучшению результатов, поскольку для сложных молекул зачастую требуется множество стадий. Кроме того, с точки зрения химической логики часто бывает так, что более длинная схема с задействованием стратегии защитных групп оказывается более реализуемой и удобной в исполнении по сравнению с короткой, но неселективной схемой. Поэтому в данном случае наиболее разумным видится использование глубины в 9 стадий.

Таким образом, в качестве простых параметров запуска «по-умолчанию» из общих соображений были выбраны время 10 минут и глубина 9 стадий. Они являются компромиссными как с точки зрения качества результатов, так и с позиций удобства взаимодействия с инструментом. Именно эти параметры были использованы для дальнейших экспериментов по варьированию стоков.

3.3 Влияние стоков

Как уже говорилось ранее, стоки имеют очень высокое влияние на качество предсказаний, поскольку, доходя именно до молекул из стоков, алгоритм прекращает дальнейшее расширение. Наличие широкого спектра реагентов может значительно улучшить качество предсказаний за счет выбора более оптимальных путей синтеза и сокращения среднего числа реакций на схему. Для создания и предподготовки стоков был написан скрипт,

основанный на использовании пакета RDKit [34] и осуществляющий проверку корректности введенных SMILES, их стандартизацию и перевод в InChiKey с последующим сохранением в файл формата hdf5 (рисунок 27).

Рисунок 27 – Процесс подготовки стоков

3.3.1 Сток с in-house данными

В рамках химического департамента компании Biocad существует собственный веб-сервис ChemSoft, использующийся для занесения и хранения данных о доступных соединениях, химических реакциях и т. д. Во внутренней библиотеке соединений на сегодняшний день насчитывается около 10 тысяч молекул, среди которых есть множество довольно удобных и широко распространенных реагентов, рассчитанных на сборку сложных соединений почти с нуля. Поэтому использование данного набора соединений в качестве стока выглядит разумным решением, которое также позволит получать более репрезентативные схемы из соединений непосредственно «на полке».

Для создания стока chemsoft из соответствующей базы данных были выгружены SMILES всех когда-либо занесенных туда соединений. После предобработки этих данных полученный сток chemsoft (сокращение – cs; 9298 молекул) был использован для анализа тестовых датасетов в нескольких условиях.

В случае базовых условий (2 минуты и 6 стадий) были получены следующие результаты: для тестового набора была решена 21 молекула, а для набора AiZynthFinder – 26 молекул, что несколько меньше, чем в случае использования стока zinc (31 и 40 молекул соответственно). При переходе же к оптимальным условиям (10 минут и 9 стадий) получились результаты, сравнимые с таковыми для стока zinc: для тестового набора и стока chemsoft – 37 молекул против 39 для стока zinc, а для набора AiZynthFinder – 40 против 46 соответственно (рисунок 28).

Схожесть числа решенных молекул в оптимальных условиях оказалась весьма неожиданной, поскольку между сравниваемыми стоками существует огромная разница с точки зрения числа входящих в них молекул: сток chemsoft насчитывает чуть больше 9 тысяч молекул, в то время как в стоке zinc хранится более 17 миллионов молекул. Вероятно, при создании данного стока из-за примененных фильтров разработчики инструмента AiZynthFinder могли пропустить большое количество полезных органических и неорганических реагентов.

Стоит отметить, что разница между стоками в количестве и природе входящих в них молекул не могла не отразиться на общем виде химических схем. Так, в случае стока chemsoft схемы имеют большее количество стадий по сравнению со стоком zinc, причем эта тенденция становится более заметной при переходе от базовых условий к оптимальным (рисунок 29). Объясняется данное поведение использованием в стоке chemsoft более простых реагентов, требующих большее количество химических превращений. Однако, несмотря на удлинение схем, они представляются более полезными, поскольку задействуют соединения из собственного каталога, а значит могут быть реализованы текущими ресурсами.

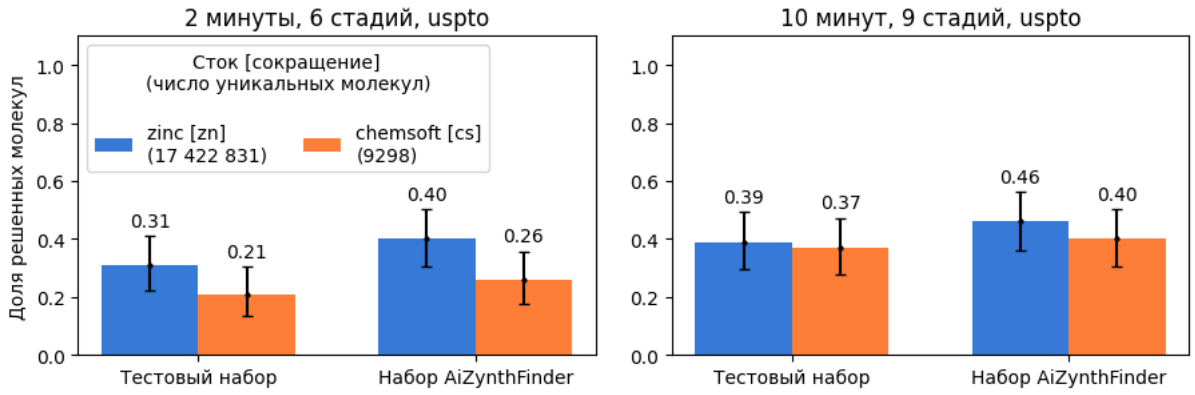

Рисунок 28 – Доля решенных молекул для стоков chemsoft и zinc

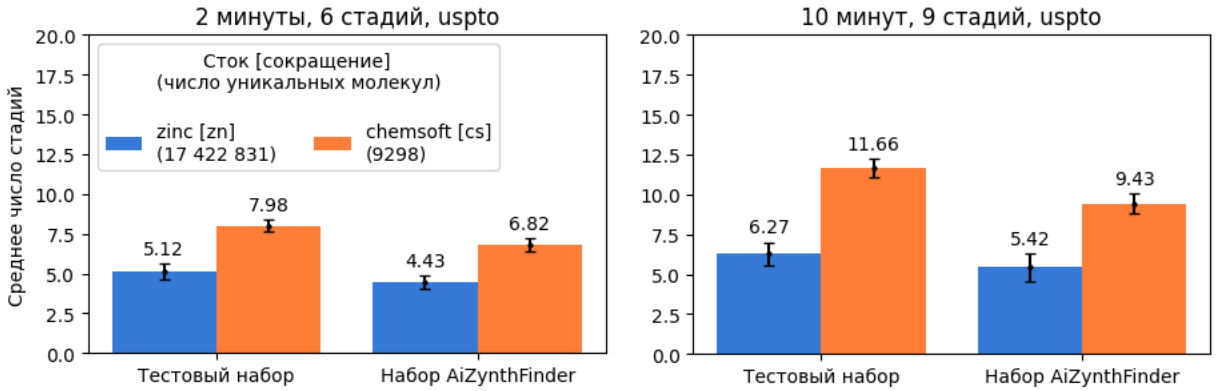

Рисунок 29 – Среднее число стадий для стоков chemsoft и zinc

Таким образом, даже на примере добавления стока chemsoft можно сделать вывод, что использование репрезентативных наборов молекул делает предсказания алгоритма более качественными и, вероятно, позволит успешно решать большее число соединений. Всё это делает актуальной задачу добавления новых стоков для их дальнейшего использования в ретросинтетическом анализе.

3.3.2 Стоки из открытых баз данных

В качестве источника информации для создания стоков перспективной оказалась база данных PubChem [53]. Во-первых, на сайте PubChem существует множество фильтров для поиска необходимых под конкретную

задачу соединений (поиск по патентам, статьям, биологической активности и т. д.). Во-вторых, с помощью данного ресурса можно получать соединения, представленные в каталогах множества поставщиков химических реагентов, что позволит ретросинтетическому инструменту ссылаться на сайты подходящих магазинов. В-третьих, на страницах конкретных соединений с сайта PubChem находится огромное количество информации о молекуле, например, химические и физические свойства, ссылки на литературу, токсичность и многое другое. Всё это поможет сделать ретросинтетический инструмент более интерактивным и информативным.

Для создания стока pubchem была выгружена выборка соединений, для которых имеется спектральная информация (1 619 150 молекул). Наличие спектров говорит о том, что молекула охарактеризована, а значит может быть использована для синтеза и получения новых соединений. После обработки сырых данных был получен сток pubchem (сокращение – pc), состоящий из 1 603 094 молекул.

Также с сайта PubChem были выгружены наборы соединений от таких поставщиков, как BLDPharm [54], Angene Chemical [55] и LeapChem [56]. Из данных наборов были получены стоки bld (1 186 862 соединений), angene (сокращение – ag, 224 110 соединений) и leapchem (сокращение – lc, 24 350 соединений).

Наконец, в дополнение к уже имеющимся наборам был создан сток chembl (сокращение – cb, 1 824 787 соединений) из малых молекул базы ChEMBL. В этой базе данных хранится множество полезных билдинг-блоков и реагентов, для которых, по аналогии с сайтом PubChem, есть страницы с большим количеством полезной информации.

На рисунке 30 представлены результаты предсказаний на тестовых наборах при использовании перечисленных выше стоков (сортировка в порядке увеличения числа молекул в стоке).

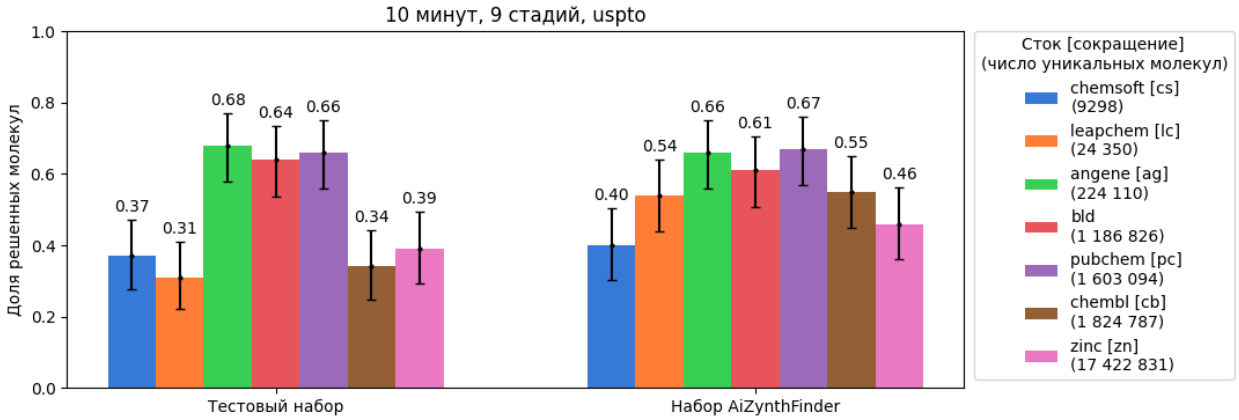

Рисунок 30 – Доля решенных молекул в зависимости от стока

Исходя из экспериментальных данных можно сделать ряд выводов:

Во-первых, для обоих тестовых наборов очевидным образом выделяется группа результатов, полученных при использовании стоков angene, bld и pubchem. В данной группе все запуски получили долю решенных молекул выше 60% (вплоть до 68% для тестового набора и стока angene), что является приростом на более чем 50% по сравнению с результатами стока chemsoft. Объяснить данное улучшение можно природой перечисленных стоков: Angene и BLD – это поставщики наиболее распространенных билдинг-блоков и реагентов, а сток PubChem содержит в себе охарактеризованные молекулы со спектрами. То есть, в данных стоках содержатся молекулы, чаще всего использующиеся для синтеза органических соединений, что позволяет алгоритму строить более репрезентативные схемы и решать большее число молекул.

Во-вторых, довольно любопытные результаты были получены в случае стоков leapchem и chembl: в рамках одного набора доли решенных молекул для данных стоков почти не отличались, в то время как при сравнении наборов между собой в долях решенных молекул обнаружилась заметная разница. Так, для тестового набора были решены 31 (leapchem) и 34 (chembl) молекулы, что оказалось даже несколько хуже результатов стоков chemsoft (37 молекул) и zinc (39 молекул). Набор AiZynthFinder в свою очередь показал более

выигрышные результаты по сравнению со стоками chemsoft и zinc – 54 (leapchem) и 55 (chembl) молекул против 40 (chemsoft) и 46 (zinc). При этом данные результаты в обоих случаях были хуже, чем у группы стоков angene, bld и pubchem. Вероятно, связано такое поведение в первую очередь с природой тестовых наборов: как уже было показано ранее, они заметно отличаются друг от друга, и набор молекул AiZynthFinder оказался более подходящим для решения с помощью стоков leapchem и chembl. Кроме того, сами стоки leapchem и chembl имеют заметные отличия от предыдущей группы: сток leapchem содержит чуть меньше 25 тысяч соединений, многие из которых относятся к вспомогательным реагентам и редко используются алгоритмом; в стоке chembl, напротив, находится множество слишком продвинутых молекул, не подходящих под понятие универсальных реагентов.

В-третьих, подтверждается обнаруженное в пункте 3.3.1 наблюдение об отсутствии зависимости числа молекул в стоке и количества решенных молекул. Сток zinc, обладающий самым большим числом молекул (> 17 млн), проигрывает значительно менее многочисленным стокам pubchem, bld (< 2 млн молекул на каждый) и даже стоку angene (< 300 тыс. молекул). В данном случае наиболее важными характеристиками стоков видятся их качество и репрезентативность входящих в них молекул, а не размер.

С точки зрения числа стадий также видим довольно отличающиеся результаты у разных стоков (рисунок 31). Лидером для обоих наборов молекул является сток bld, демонстрирующий самые короткие схемы. Сток chemsoft в свою очередь остается самым «жадным» до числа стадий из-за уже упоминавшейся природы входящих в него соединений.

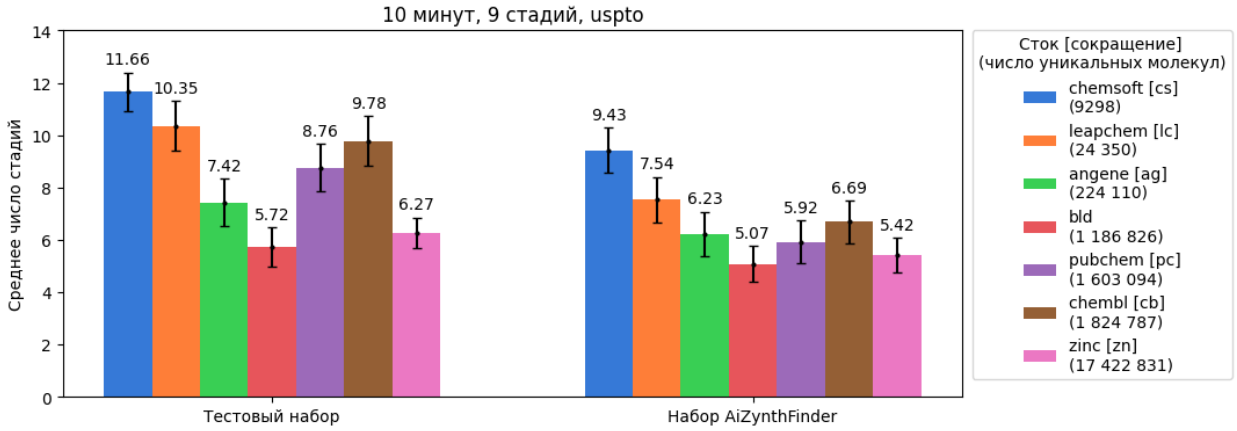

Рисунок 31 – Среднее число стадий в зависимости от стока

Таким образом, с помощью использования более репрезентативных стоков удалось добиться увеличения доли решенных молекул до 67–68% и повышения качества получаемых схем. Однако, результаты можно улучшить ещё сильнее, если использовать комбинации стоков.

3.3.3 Использование комбинаций стоков

В рамках инструмента AiZynthFinder существует возможность использования в одном анализе сразу нескольких стоков. Поскольку каждый из рассмотренных ранее стоков имеет собственное происхождение, наборы входящих в них молекул также имеют очень разную природу. Благодаря использованию различных комбинаций стоков можно взаимно скомпенсировать их недостатки и покрыть значительно более широкую область химического пространства, чем при использовании стоков по отдельности.

На рисунке 32 представлены результаты экспериментов по использованию различных комбинаций стоков. Для того, чтобы сократить количество запусков инструмента, рассматривалась только линейка экспериментов с последовательным добавлением всё более объемных наборов молекул.

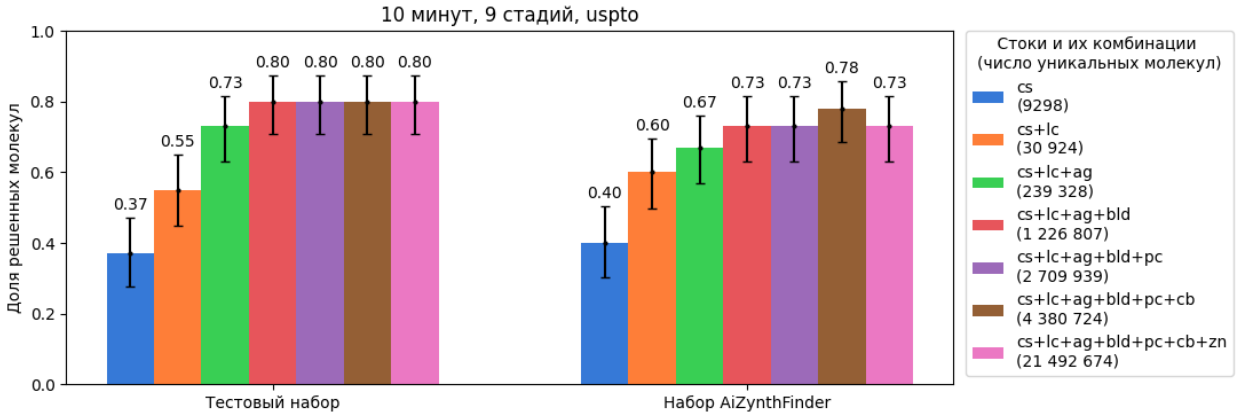

Рисунок 32 – Доля решенных молекул в зависимости от комбинации стоков

Как и ожидалось, доля решенных молекул увеличивается по мере добавления в анализ новых стоков. Интересно, что в случае обоих наборов, начиная с комбинации «cs+lc+ag+bld» (1 266 807 уникальных молекул) наблюдается выход на своеобразное плато результатов: 80% для тестового набора и 73% для набора AiZynthFinder. Последующее увеличение числа уникальных молекул в комбинации стоков почти не влияет на итоговые цифры. Исключение составляет лишь случай использования предпоследней комбинации стоков (все стоки, кроме zinc; 4 380 724 молекулы) для набора AiZynthFinder: здесь мы видим 78 решенных молекул из 100, что на 5 больше, чем у комбинаций по соседству. Возможно, такой выброс может быть связан с тем, что в стоке zinc были найдены прекурсоры, которые повели поиск по дереву в неверном направлении с точки зрения общего качества итоговой схемы. В отсутствии данных прекурсоров алгоритм строил схемы иначе, что позволило решить на пять молекул больше. Тем не менее, говорить о статистической значимости данной закономерности не приходится, особенно с учетом ровного плато у тестового набора.

Среднее число реакций при добавлении новых стоков, напротив, уменьшалось, выходя на плато в районе 5 стадий. Данное наблюдение справедливо для обоих наборов (рисунок 33).

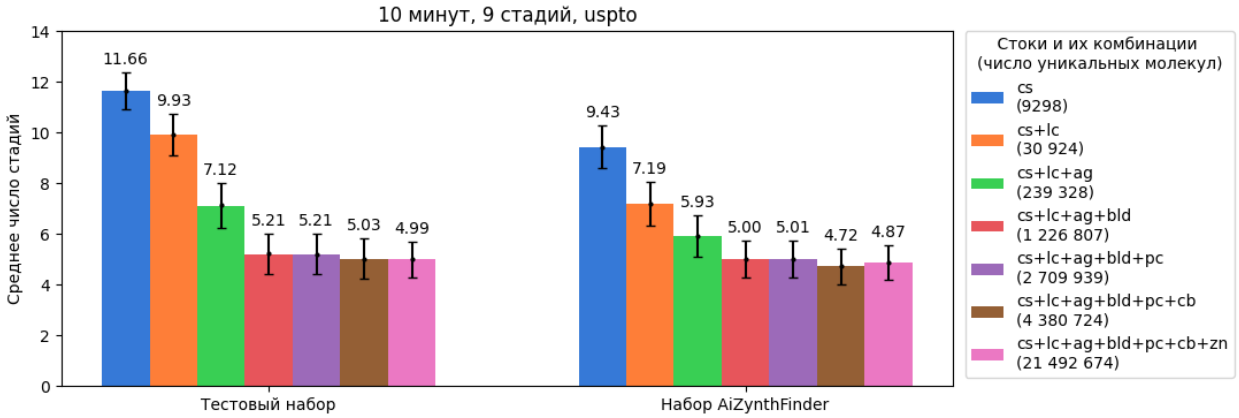

Рисунок 33 – Среднее число стадий в зависимости от комбинации стоков

Дополнительно были поставлены эксперименты с комбинациями стоков, не входящими в исследованный ранее набор входных условий (рисунок 34): «cs+zn» – комбинация самого маленького и самого большого стока (17 428 094 уникальных молекул); «ag+bld+pc» – комбинация самых объемных и репрезентативных стоков (2 704 579 уникальных молекул).

Интересно, что для тестового набора использование комбинации стоков chemsoft и zinc позволило увеличить долю решенных молекул более чем на 50% (с 37–39% по отдельности до 63% вместе). Для набора AiZynthFinder синергический эффект комбинации стоков тоже наблюдался, но в меньшей степени.

Также любопытным кажется тот факт, что использование набора стоков «ag+bld+pc» позволило приблизиться (для тестового набора) и даже несколько превзойти (для набора AiZynthFinder) результаты плато в рассмотренных выше экспериментах (см. рисунок 32).

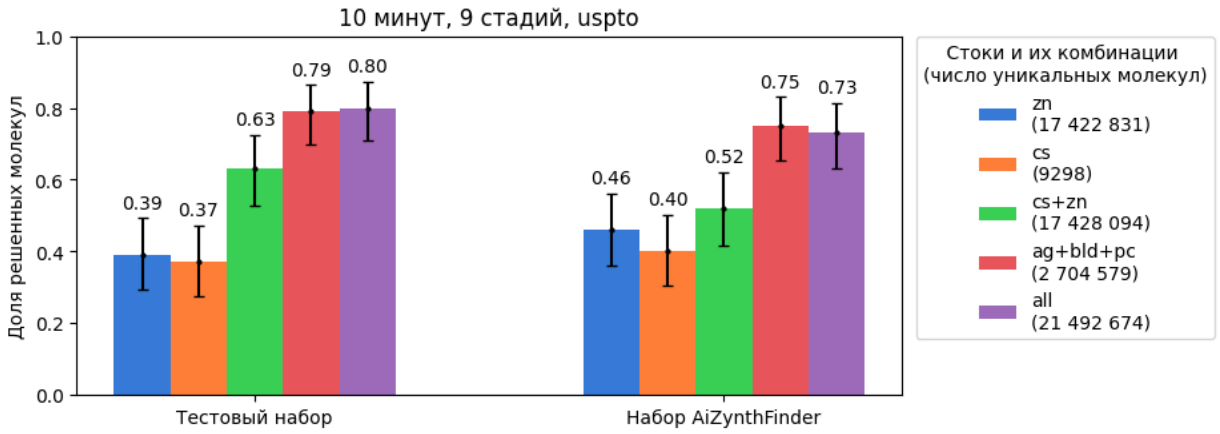

Рисунок 34 – Дополнительные эксперименты с комбинациями стоков

Таким образом, с использованием комбинаций стоков удалось добиться значительного прироста доли решенных молекул в обоих тестируемых наборах (вплоть до 75–80%).

3.3.4 Выводы по стокам

Как было показано в рассмотренных ранее пунктах, стоки оказывают огромное влияние на качество предсказаний. Это выражается как в количестве успешно решаемых молекул (на тестовых наборах были достигнуты значения вплоть до 80% решенных молекул), так и в качестве самих ретросинтетических схем (уменьшается число стадий, схемы становятся более «химичными», наблюдается меньшее число галлюцинаций). Возможность использования различных комбинаций стоков позволяет направить схемы в нужное пользователю русло, в том числе с помощью ограничения ряда используемых стоков: например, можно использовать только стоки из in-house библиотеки или только стоки от поставщиков. Особенно удобно, что по результатам анализа можно получить выжимку информации о билдинг-блоках, содержащихся в конкретных стоках. Всё это делает рассматриваемый ретросинтетический инструмент своего рода «умным» помощником в поиске подходящих реагентов для синтеза целевых молекул.

3.4 Влияние стратегий расширения

В пункте 2.1.1 упоминалось, что в рамках стандартного пакета AiZynthFinder версии 3.6.0 существует дополнительная стратегия расширения ringbreaker, направленная на разбиение циклов. Как самостоятельная стратегия она представляется не самой полезной, поскольку строит неполные схемы, однако вместе со стандартной стратегией uspto, гипотетически, может положительно сказаться на количестве решенных молекул.

На рисунке 35 представлены результаты тестирования различных комбинаций стоков в условиях использования комбинации стратегий расширения «uspto+ringbreaker».

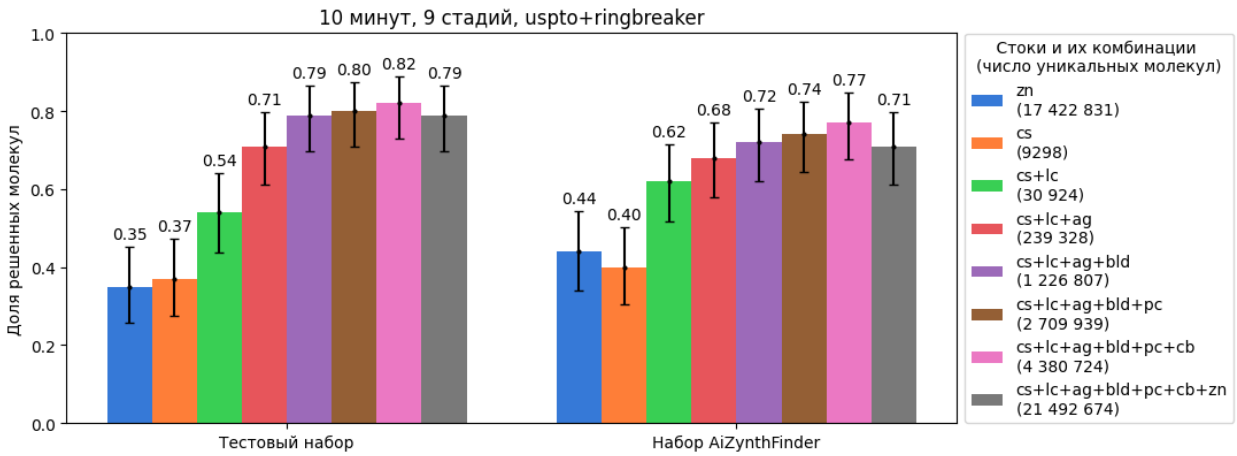

Рисунок 35 – Результаты тестирования комбинации стратегий расширения

При рассмотрении результатов и сравнении их со стандартной стратегией расширения uspto (рисунок 32) напрашивается вывод, что использование стратегии ringbreaker в комбинации с uspto не принесло существенных результатов. В большинстве случаев были получены либо идентичные, либо чуть меньшие доли решенных молекул. Интересно, что в этом эксперименте использование комбинации стоков без zinc снова дало чуть более высокий результат по сравнению с использованием всех стоков. Это

несколько укрепляет уверенность в гипотезе о наличии в стоке цинка прекурсоров, ведущих к ухудшению качества итоговых схем.

Таким образом, использование комбинации стратегий расширения не сильно повлияло на итоговое качество предсказаний. Связано это, вероятно, с тем, что стратегия разбиения ringbreaker встречается в схемах крайне редко (выбор в основном отдается uspto), поэтому её влияние на схемы оказывается минимальным. Использовать в анализе данные условия можно, но стоит понимать, что предсказания могут почти не измениться или даже ухудшиться.

3.5 Валидация результатов

Для валидации результатов было решено провести тестирование инструмента в наиболее интересных условиях на выборке в 1000 молекул. Выборка была получена с помощью уже упомянутых ранее фильтров (рисунок 16) из соединений базы ChEMBL, находящихся уже на 3 фазе клинических испытаний. Молекулы анализировались с помощью стратегии расширения uspto в следующих условиях:

1) Сток zinc, 2 минуты, 6 стадий;
2) Все стоки, 10 минут, 9 стадий;
3) Все стоки без zinc, 10 минут, 9 стадий.

Результаты полученных экспериментов приведены на рисунке 36. В базовых условиях (№1) было решено 377 молекул, в условиях №2 – 793 молекулы, а в условиях №3 – 810 молекул. Видим, что переход к более объемному тестовому набору почти не повлиял на доли решенных молекул: похожие результаты были получены и ранее на двух наборах по 100 молекул.

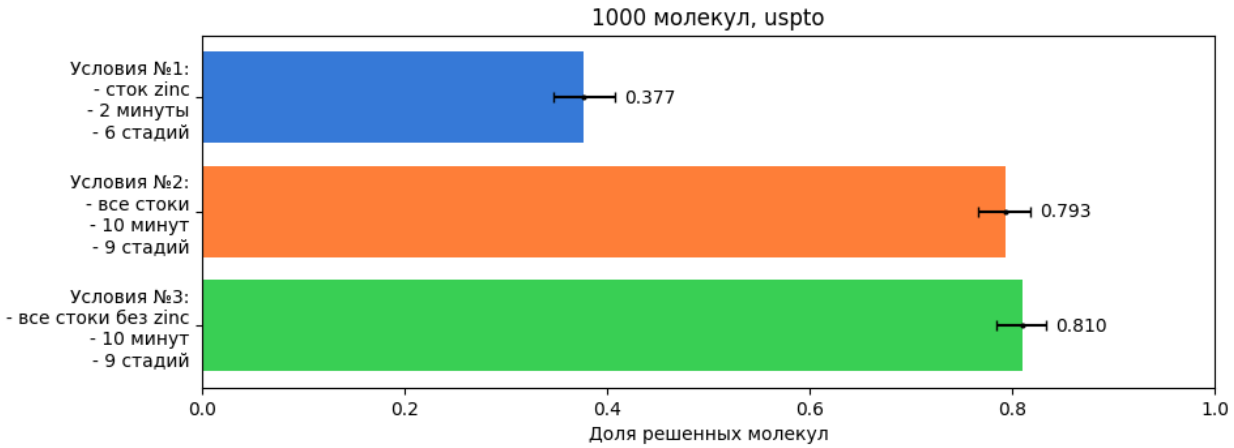

Рисунок 36 – Результаты экспериментов на наборе из 1000 молекул

Интересно, что тенденция некоторого улучшения результатов при переходе от условий №2 к условиям №3 сохраняется и здесь. Чтобы с уверенностью утверждать отличие данных условий, необходимо использовать кратно большую тестовую выборку. Тем не менее, от стока zinc как стока «по-умолчанию» можно отказаться по следующим причинам:

– Во-первых, наполнение данного стока является довольно абстрактным для конечного пользователя. Авторы инструмента AiZynthFinder применили достаточно жесткие фильтры при формировании стока zinc, отчего туда не попало множество полезных реагентов;

– Во-вторых, данный сток имеет самый низкий коэффициент полезного действия (отношение числа решенных молекул к числу молекул в стоке) среди всех протестированных наборов. С помощью стока zinc (включает в себя более 17 миллионов молекул) решается примерно та же доля тестовых наборов, что и в случае стока chemsoft (включает в себя чуть больше 9 тысяч молекул);

– В-третьих, из-за огромного количества входящих в состав стока zinc молекул, его загрузка в оперативную память при запуске ретросинтетической задачи занимает больше всего времени;

– В-четвертых, при формировании данного стока авторы инструмента не оставили ссылок на входящие в него молекулы, что лишает инструмент части интерактивности при использовании данного стока (см. подраздел 4.2 и рисунок 43).

Кроме того, преимущество комбинации стоков без zinc транслируется для большинства протестированных условий (см. рисунок 39). Все эти факторы делают использование условий №3 с комбинацией стоков без zinc более предпочтительными по сравнению с условиями №2. Далее будем считать условия №3 оптимальными.

Рассмотрим другие характеристики результатов, полученных для большой тестовой выборки. На рисунке 37 представлено сравнение синтетической доступности решенных и нерешенных алгоритмом молекул в различных условиях. Переход от базовых условий (№1) к оптимальным (№3) позволил решать более широкий диапазон молекул с точки зрения их сложности. Те же молекулы, для которых решение найдено не было, также становятся более сложными. Таким образом, в оптимальных условиях сложность молекул, которые алгоритм способен успешно решить, повышается по сравнению с базовыми условиями.

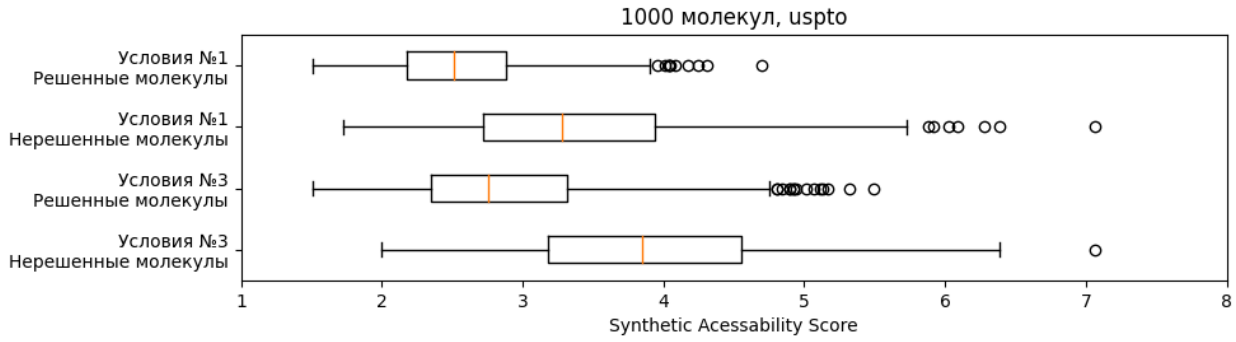

Рисунок 37 – Сравнение SAscore для решенных и нерешенных молекул

Кроме того, при переходе к оптимальным условиям улучшается и такая характеристика схем, как state score (см. пункт 2.1.2, RouteCollection). На

рисунке 38 видно, что в случае условий №3 значение данной метрики для большинства схем близко к единице, тогда как для условий №1 многие схемы имеют state score меньше 0.8. Это говорит о том, что при переходе к условиям №3 алгоритм строит значительно больше удачных схем.

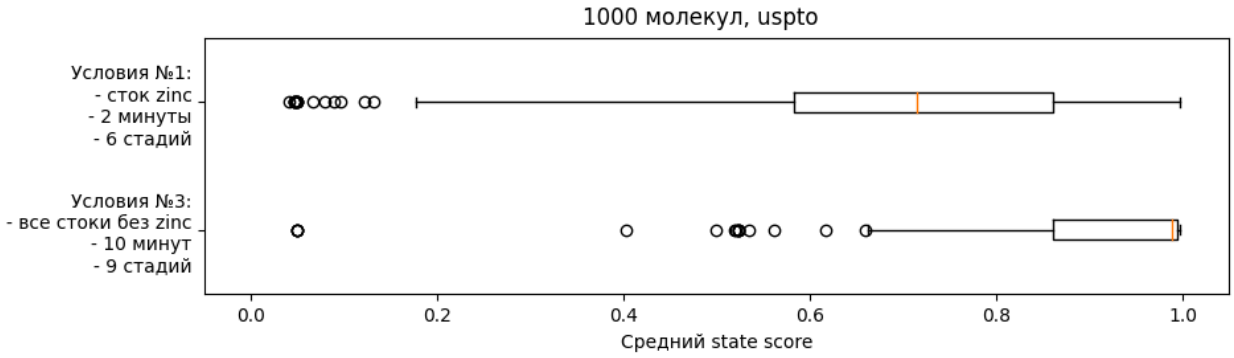

Рисунок 38 – Сравнение средних значений state score

Таким образом, переход от базовых условий №1 к оптимальным условиям №3 позволил улучшить практически все характеристики получаемых результатов.

3.6 Вывод

На рисунке 39 представлены ключевые результаты работы. В рамках проведенного исследования были протестированы все ключевые параметры запуска инструмента AiZynthFinder. В ходе тестирования инструмента было поставлено более 13 тысяч задач на ретросинтетический анализ для 1200 уникальных молекул. Если считать, что среднее время выполнения одного анализа – 10 минут (на самом деле оно больше), то последовательный беспрерывный запуск всех проведенных анализов занял бы более 3 месяцев. Благодаря возможности запускать вычисления на кластере параллельно этот процесс занял значительно меньше времени.

chemsoft								✅	✅	✅	✅	✅	✅	✅	✅	
leapchem							✅		✅	✅	✅	✅	✅	✅		
angene						✅				✅	✅	✅	✅	✅		✅
bld					✅						✅	✅	✅	✅		✅
pubchem				✅								✅	✅	✅		✅
chembl			✅										✅	✅		
zinc		✅												✅	✅	
uspto 2 минуты 6 стадий	Тестовый	31	-	-	-	-	-	21	-	-	-	-	74	75	-	-
	AiZynthFinder	40	-	-	-	-	-	26	-	-	-	-	69	67	-	-
	1000 молекул	38	-	-	-	-	-	-	-	-	-	-	-	-	-	-
uspto 10 минут 6 стадий	Тестовый	41	-	-	-	-	-	27	-	-	-	-	78	78	-	-
	AiZynthFinder	46	-	-	-	-	-	38	-	-	-	-	75	73	-	-
uspto 10 минут 9 стадий	Тестовый	39	34	66	64	68	31	37	55	73	80	80	80	80	63	79
	AiZynthFinder	46	55	67	61	66	50	40	60	67	73	73	78	73	52	75
	1000 молекул	-	-	-	-	-	-	-	-	-	-	-	81	79	-	-
uspto + rb 10 минут 9 стадий	Тестовый	35	-	-	-	-	-	37	54	71	79	80	82	79	-	-
	AiZynthFinder	44	-	-	-	-	-	40	62	68	72	74	77	71	-	-

Рисунок 39 – Доля решенных молекул (%) в различных условиях

Было показано, что наибольшее влияние на качество итоговых предсказаний оказывает использование репрезентативных стоков и их комбинаций. Несмотря на то, что идущий в комплекте с AiZynthFinder сток zinc насчитывает более 17 миллионов соединений, он решает почти то же число молекул, что и сток chemsoft, в котором содержится чуть меньше 10 тысяч реагентов. То есть размер стока почти не влияет на результаты в отличие от его наполнения. Благодаря инжинирингу данных и подбору подходящих стоков из открытых источников удалось увеличить долю решаемых алгоритмом молекул вплоть до 80%. Полученные результаты подтверждаются в том числе на объемной тестовой выборке в 1000 соединений.

С точки зрения простых параметров анализа, таких как время анализа и глубина схем, выбранные в качестве стандартных 10 минут и 9 стадий кажутся

наиболее универсальными. За 10 минут алгоритм, как правило, успевает найти успешные схемы, при этом выдавая достаточно стабильные результаты от запуска к запуску, чего нельзя сказать о 2 минутах. В случае глубины влияние на число решенных молекул не столь заметное, однако оно увеличивается с уменьшением количества прекурсоров в стоках, поскольку алгоритму требуется больше стадий для нахождения удачных схем. Если использовать в комбинации много стоков, то схемы обычно решаются раньше, чем достигают ограничения по глубине.

Кроме влияния на долю решенных молекул, переход к оптимальным условиям также привел и к улучшению других параметров получаемых схем синтеза. Во-первых, расширился диапазон синтетической доступности молекул, способных быть решенными инструментом: такие молекулы в среднем стали сложнее. Во-вторых, алгоритм начал выдавать значительно больше удачных схем, о чем говорит близкий к единице средний state score. В-третьих, алгоритм начал выдавать схемы с меньшим количеством стадий, что делает многие из них более применимыми в реальной синтетической практике.

Таким образом, оптимальными параметрами запуска алгоритма видятся следующие: время анализа – 10 минут, глубина – 9 стадий, стоки – все доступные кроме стока zinc, стратегия расширения – uspto. В зависимости от молекулы и целей пользователя данные параметры могут варьироваться, однако первичный анализ следует проводит в таких условиях. Перечисленные параметры были выбраны в качестве условий «по-умолчанию» для ретросинтетического инструмента, внедренного в вычислительный веб-сервис компании Biocad. Более подробно об этом написано в следующей главе.

4 Интеграция инструмента

4.1 Архитектура

ChemLab – это большой вычислительный веб-сервис химического департамента компании Biocad, включающий в себя множество подсервисов, таких как предсказание метаболизма, комбинаторная генерация молекул и т.д. В рамках данной работы было необходимо внедрить в ChemLab ретросинтетический инструмент на основе AiZynthFinder. Для этого была разработана cli-утилита (command line interface, интерфейс командной строки) для постановки ретросинтетических задач на вычислительном кластере компании. С помощью данной утилиты можно запускать задачу на анализ одной молекулы с перечисленными в пункте 2.1.3 параметрами AiZynthFinder и получать модифицированные результаты анализа вместе с расширенной статистикой. На рисунке 40 представлена примерная схема работы ChemLab при запуске ретросинтетического предсказания.

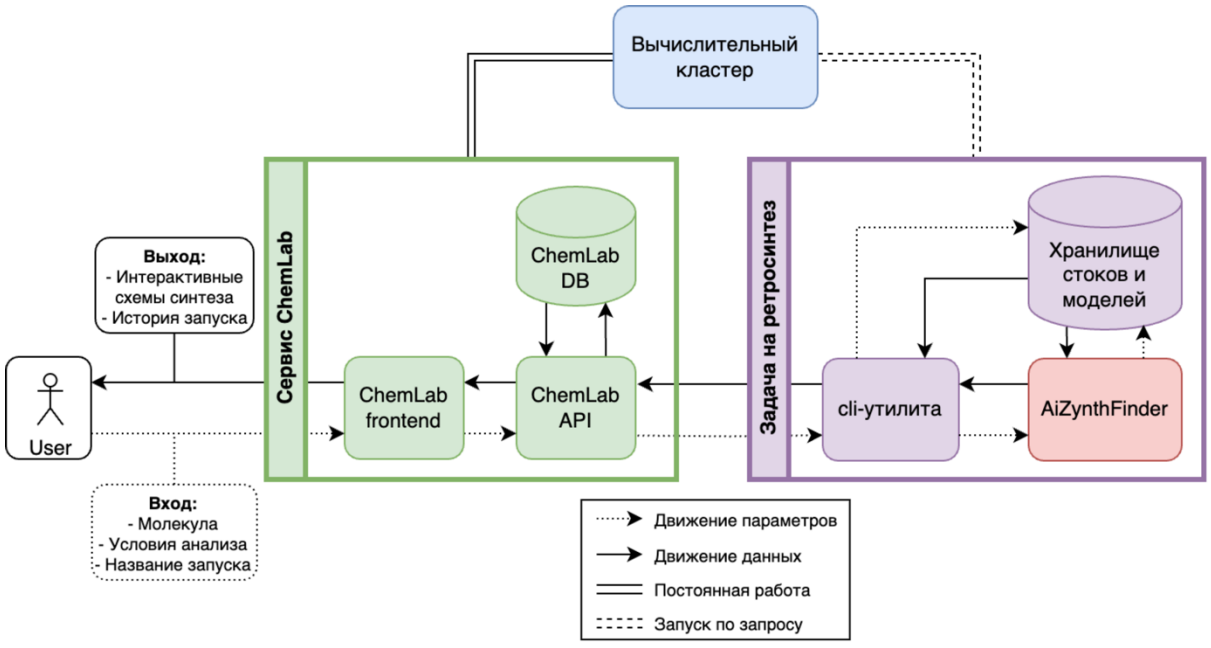

Рисунок 40 – Схема работы ретросинтеза в ChemLab

В веб-интерфейсе (подраздел 4.2) пользователь ставит ретросинтетическую задачу, задавая необходимые параметры анализа. Задача обрабатывается с помощью ChemLab API, который обращается к вычислительному кластеру и передает параметры в cli-утилиту. В рамках работы cli-утилиты происходит проверка корректности параметров и наличия критических файлов в подключаемом к задаче хранилище на кластере. Затем инициализируется AiZynthFinder, он загружает файлы конфигурации, стоков и моделей и производит ретросинтетический анализ. По окончании анализа его результаты дополнительно обрабатываются cli-утилитой и возвращаются в ChemLab API для сохранения в базе данных ChemLab и удобного вывода пользователям.

4.2 Пользовательский интерфейс

В ChemLab был разработан удобный пользовательский интерфейс для постановки ретросинтетических задач (рисунок 41). В рамках данного интерфейса пользователь выбирает время и глубину анализа (либо оставляет параметры по умолчанию), вводит целевую молекулу удобным ему способом (рисование, ввод SMILES, прикрепление файла), выбирает стоки и стратегию расширения (либо оставляет параметры по умолчанию), дает название анализу и отправляет задачу на вычисление. Также присутствует возможность отправить на вычисление сразу несколько молекул.

После постановки задачи пользователь переходит в историю запусков инструмента (рисунок 42), где можно отслеживать статус выполнения задач и переходить на страницу с результатами анализа.

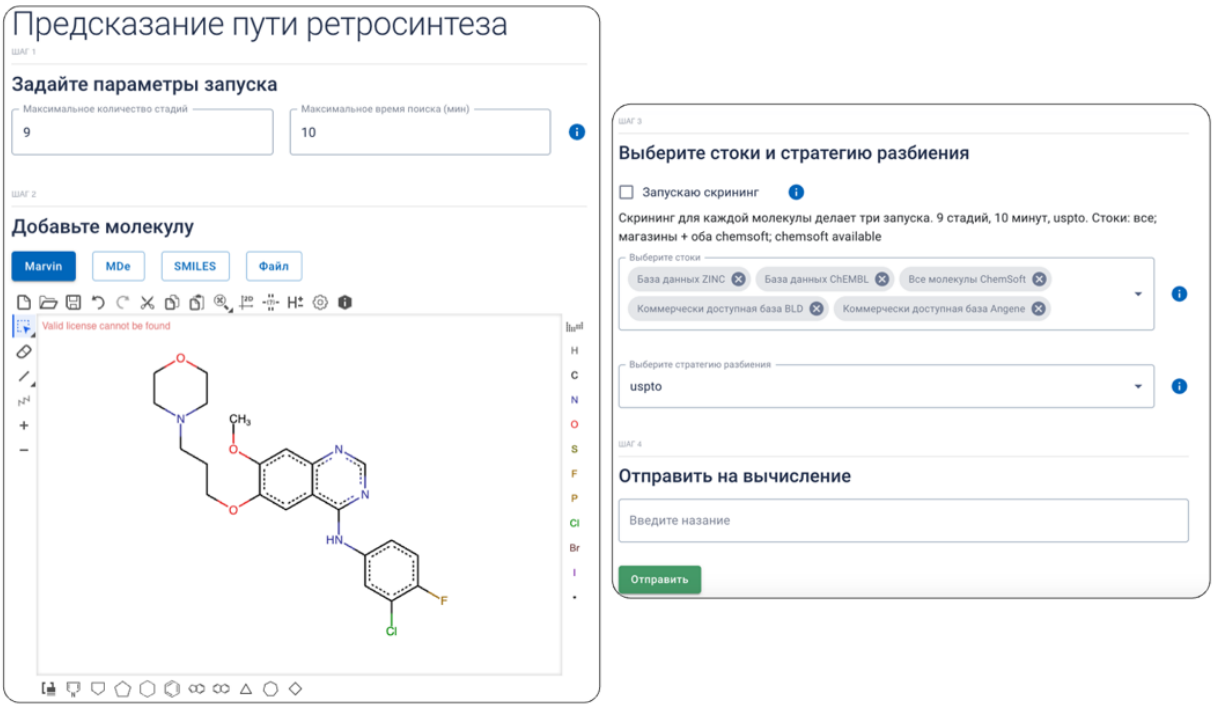

Рисунок 41 – Интерфейс постановки ретросинтетической задачи в ChemLab

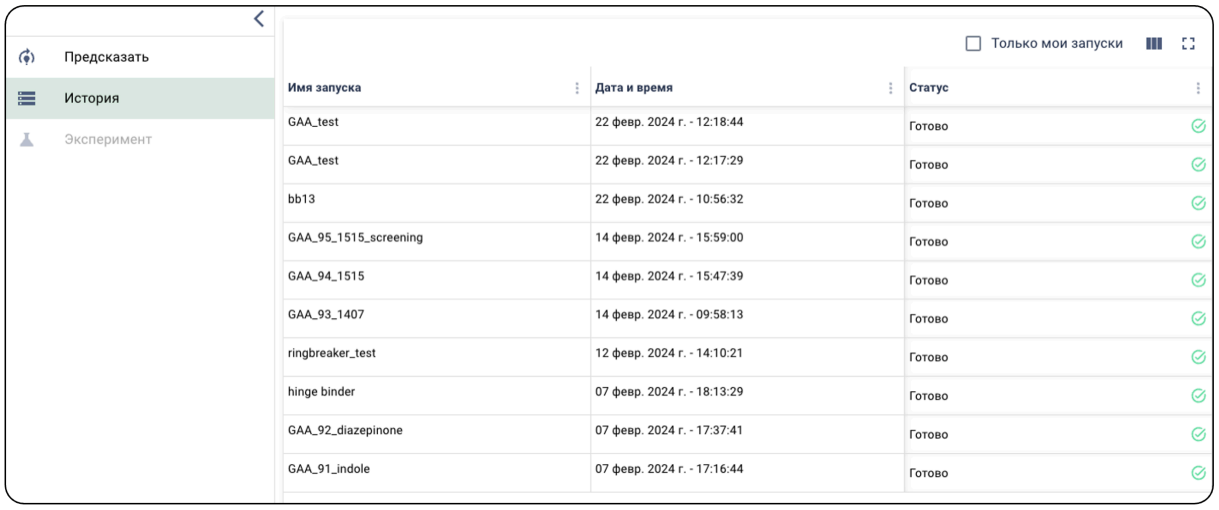

Рисунок 42 – Интерфейс истории запусков ретросинтеза в ChemLab

Сами результаты отображаются в ChemLab в виде интерактивных схем синтеза, позволяющих увеличивать изображения молекул, копировать их SMILES, просматривать информацию о наличии в стоках для доступных молекул и переходить по ссылкам на страницы конкретных прекурсоров

(рисунок 43). Кроме того, пользователь может удобно переключаться между схемами и просматривать статистику анализа.

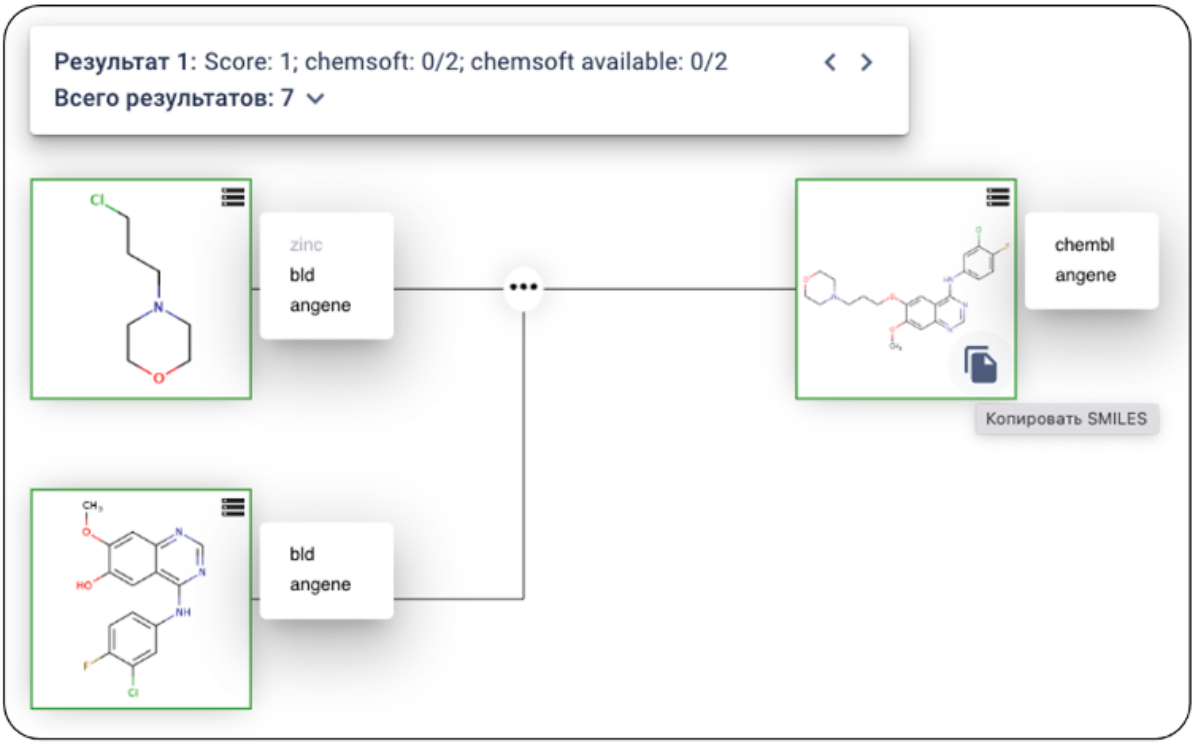

Рисунок 43 – Вид ретросинтетических схем в ChemLab

Всё это делает взаимодействие с результатами ретросинтетического анализа значительно более удобным и наглядным для пользователей.

4.3 Технические особенности

Cli-утилита была написана на языке программирования Python версии 3.9, ключевыми программными пакетами являются AiZynthFinder [37], Click [57], Loguru [58], RDKit [34].

Контейнеризация осуществляется с помощью Docker [59], CI/CD – с помощью GitLab [60] и Kubernetes [61]. Для постановки одной ретросинтетической задачи используется 16 гигабайтов RAM и одно ядро

процессора, операционная система – Ubuntu 22.04 [62]. GPU для анализа не требуется.

ChemLab API реализован с помощью Python FastAPI [63], ChemLab frontend написан на языке программирования TypeScript [64].

4.4 Вывод

На основании пакета AiZynthFinder был разработан собственный веб-сервис с удобным пользовательским интерфейсом. Данный сервис позволяет запускать ретросинтетический анализ с необходимыми параметрами, отображать его результаты в виде интерактивных схем синтеза, а также хранить историю запусков инструмента. Оптимальные параметры запуска, полученные в рамках описанного ранее исследования (глава 3), используются в качестве параметров по-умолчанию.

ЗАКЛЮЧЕНИЕ

В данной дипломной работе была поставлена основная цель – проведение исследования факторов, влияющих на качество предсказаний ретросинтетических схем с помощью инструмента AiZynthFinder. На основании исследования было необходимо подобрать оптимальные параметры по-умолчанию для запуска предсказаний в собственном вычислительном веб-сервисе. Для достижения цели был выполнен ряд задач.

Первым этапом работы стало рассмотрение существующих подходов к одностадийному и многостадийному ретросинтетическому планированию. Были описаны ключевые способы решения этих задач, среди которых наиболее эффективным и быстроразвивающимся оказался подход, основанный на использовании нейросетевых моделей. Проведенный анализ моделей для многостадийного ретросинтетического планирования показал, что наиболее удобным для интеграции в собственный веб-сервис является инструмент AiZynthFinder, имеющий открытый исходный код и удобный python-интерфейс. Также для данного инструмента были выявлены аспекты, требующие доработки, а именно подбор оптимальных условий запуска и репрезентативных стоков.

Во второй главе дипломной работы была описана архитектура AiZynthFinder, ввод и вывод инструмента, а также ключевые параметры его запуска. Кроме того, была сформулирована методология оценки качества результатов ретросинтетического анализа, сформированы тестовые датасеты (2 набора по 100 молекул), определены ключевые метрики качества ретросинтетических предсказаний.

В третьей главе представлены результаты исследования факторов, влияющих на итоговый вид ретросинтетических предсказаний инструмента. Благодаря подбору подходящих параметров запуска и репрезентативных стоков удалось увеличить долю решаемых молекул с 30–40% (для базовых

условий) до 70–80% (для оптимальных условий). Валидация результатов на объемной выборке в 1000 молекул показала аналогичные метрики, что подтверждает универсальность подобранных условий запуска ретросинтетического анализа.

Четвертая глава описывает интеграцию ретросинтетической модели в собственный вычислительный веб-сервис. На основании python-интерфейса AiZynthFinder была написана cli-утилита для постановки ретросинтетических задач на вычислительном кластере компании Biocad. С помощью этой утилиты производится запуск предсказаний и обработка их результатов для последующего вывода интерактивных химических схем в рамках веб-сервиса ChemLab. Также организовано хранение истории обращений к инструменту и удобный интерфейс для постановки задач. Полученные в рамках третьей главы оптимальные параметры запуска выставлены в сервисе ChemLab в качестве параметров по-умолчанию.

Таким образом, в ходе дипломной работы на основании инструмента AiZynthFinder был создан собственный ретросинтетический сервис с удобным пользовательским интерфейсом. Данный сервис активно используется химиками-синтетиками и продуктовыми командами компании Biocad, позволяя значительно ускорить разработку лекарственных препаратов на основе малых молекул. Благодаря подбору оптимальных параметров запуска и репрезентативных входных данных доля успешно решаемых моделью молекул была увеличена до 80%.

Исследование влияния стоков и простых параметров анализа на качество предсказания ретросинтетических схем было проведено впервые и является существенным вкладом в развитие автоматизированного ретросинтетического анализа. Ранее в литературе упоминалось лишь влияние обучающих датасетов на различные характеристики предсказания одностадийной ретросинтетической реакции [65]. Полученные в рамках работы результаты были опубликованы в сборнике тезисов Всероссийской научной студенческой

конференции «ИНТЕР – Информационные технологии и радиоэлектроника 2024» (ISBN: 978-5-91256-646-2).

В заключении хотелось бы отметить, что данная работа является лишь начальным этапом усовершенствования представленного инструмента. Перспективными направлениями для последующих улучшений видятся следующие аспекты:

– Создание собственной ретросинтетической модели на основании in-house данных о химических реакциях, проведенных в компании Biocad;

– Создание новых моделей на основе открытых баз данных, таких как Open Reaction Database [66];

– Внедрение в алгоритм поиска по дереву Монте-Карло более продвинутых моделей для одностадийного ретросинтетического анализа на основе нейросетей.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Sarker, I. H. Machine Learning: Algorithms, Real-World Applications and Research Directions / I. H. Sarker // SN Computer Science. – 2021. – Vol. 2. – № 3. – P. 160.

2. Corey, E. J. Logic of chemical synthesis / E. J. Corey, X. M. Cheng; New ed. – New York: Wiley, 1995. – 436 p. – ISBN 978-0-471-50979-0.

3. Corey, E. J. Computer-Assisted Design of Complex Organic Syntheses / E. J. Corey, W. T. Wipke // Science. – 1969. – Vol.166. №3902. – P. 178–192.

4. Chen, J. H. No Electron Left Behind: A Rule-Based Expert System to Predict Chemical Reactions and Reaction Mechanisms / J. H. Chen, P. Baldi // Journal of Chemical Information and Modeling. – 2009. – Vol.49. №9. – P. 2034–2043.

5. Computational planning of the synthesis of complex natural products / B. Mikulak-Klucznik, P. Gołębiowska, A. A. Bayly [et al.] // Nature. – 2020. – Vol.588. №7836. – P. 83–88.

6. Retrosynthetic Reaction Prediction Using Neural Sequence-to-Sequence Models / B. Liu, B. Ramsundar, P. Kawthekar [et al.] // ACS Central Science. – 2017. – Vol.3. №10. – P. 1103–1113.

7. Coley, C. W. Computer-Assisted Retrosynthesis Based on Molecular Similarity / C. W. Coley, L. Rogers, W. H. Green, K. F. Jensen // ACS Central Science. – 2017. – Vol.3. №12. – P. 1237–1245.

8. Computer-Assisted Synthetic Planning: The End of the Beginning / S. Szymkuć, E. P. Gajewska, T. Klucznik [et al.] // Angewandte Chemie International Edition. – 2016. – Vol.55. №20. – P. 5904–5937.

9. Artificial Intelligence for Retrosynthesis Prediction / Y. Jiang, Y. Yu, M. Kong [et al.] // Engineering. – 2023. – Vol.25. – P. 32–50.

10. Recent advances in artificial intelligence for retrosynthesis / Z. Zhong, J. Song, Z. Feng [et al.] // Arxiv. – 2023. – URL: https://arxiv.org/abs/2301.05864.

11. A Unified View of Deep Learning for Reaction and Retrosynthesis Prediction: Current Status and Future Challenges / Z. Meng, P. Zhao, Y. Yu, I. King // Arxiv. – 2023. – URL: https://arxiv.org/abs/2306.15890.

12. CAS SciFinder | CAS. – URL: https://www.cas.org/cas-scifinder-discovery-platform/cas-scifindern (дата обращения: 21.01.2024) – Текст: электронный.

13. Reaxys. – URL: https://www.reaxys.com/#/login (дата обращения: 21.01.2024) – Текст: электронный.

14. Reaxys Predictive Retrosynthesis. – URL: https://www.elsevier.com/products/reaxys/predictive-retrosynthesis (дата обращения: 21.01.2024) – Текст: электронный.

15. CAS SciFinder^{n} — Retrosynthesis Software | CAS | CAS. – URL: https://www.cas.org/solutions/cas-scifinder-discovery-platform/cas-scifinder/synthesis-planning (дата обращения: 21.01.2024) – Текст: электронный.

16. SYNTHIA^{TM}. – URL: https://www.synthiaonline.com/ (дата обращения: 23.01.2024) – Текст: электронный.

17. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules / D. Weininger // Journal of Chemical Information and Computer Sciences. – 1988. – Vol.28. №1. – P. 31–36.

18. InChI, the IUPAC International Chemical Identifier / S. R. Heller, A. McNaught, I. Pletnev [et al.] // Journal of Cheminformatics. – 2015. – Vol.7. №23 – С. 1–34.

19. Machine Learning Guided Atom Mapping of Metabolic Reactions / E. E. Litsa, M. I. Pena, M. Moll [et al.] // Journal of Chemical Information and Modeling. – 2019. – Vol.59. №3. – P. 1121–1135.

20. Chen, W. L. Automatic reaction mapping and reaction center detection / W. L. Chen, D. Z. Chen, K. T. Taylor // WIREs Computational Molecular Science. – 2013. – Vol.3. №6. – С. 560–593.

21. A New Synthesis of Gefitinib / T. S. Maskrey, T. K. Matthew, G. LaPorte

[et al.] // Synlett. – 2019. – Vol.30. №04. – P. 471–476.

22. Segler, M. H. S. Planning chemical syntheses with deep neural networks and symbolic AI / M.H.S. Segler, M. Preuss, M.P. Waller // Nature. – 2018. – Vol.555. №7698. – P. 604–610.

23. Segler, M. H. S. Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction / M.H.S. Segler, M.P. Waller // Chemistry – A European Journal. – 2017. – Vol.23. №25. – P. 5966–5971.

24. Toxicity Prediction using Deep Learning / T. Unterthiner, A. Mayr, G. Klambauer, S. Hochreiter // Arxiv. – 2015. – URL: https://arxiv.org/abs/2306.15890.

25. Clevert, D. A. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs) / D. A. Clevert, T. Unterthiner, S. Hochreiter // Arxiv. – 2016. – URL: http://arxiv.org/abs/1511.07289.

26. Discovery and Preclinical Characterization of 6-Chloro-5-[4-(1-hydroxycyclobutyl)phenyl]-1H-indole-3-carboxylic Acid (PF-06409577), a Direct Activator of Adenosine Monophosphate-activated Protein Kinase (AMPK), for the Potential Treatment of Diabetic Nephropathy / K. O. Cameron, D. W. Kung, A. S. Kalgutkar [et al.] // Journal of Medicinal Chemistry. – 2016. – Vol.59. №17. – P. 8068–8081.

27. A Survey of Monte Carlo Tree Search Methods / C. Browne, E. Powley, D. Whitehouse [et al.] // IEEE Transactions on Computational Intelligence and AI in Games. – 2012. – Vol.4. №1. – P. 1–43.

28. Coulom, R. Computing Elo Ratings of Move Patterns in the Game of Go / R. Coulom // ICGA Journal. – 2007. – Vol.30. №4. – P. 198–208.

29. Move Evaluation in Go Using Deep Convolutional Neural Networks / C. J. Maddison, A. Huang, I. Sutskever, D. Silver // Arxiv. – 2015. – URL: http://arxiv.org/abs/1511.07289.

30. Lubosch, M. Industrial scheduling with Monte Carlo tree search and machine learning / M. Lubosch, M. Kunath, H. Winkler // Procedia CIRP. – 2018. –

T.72. – P. 1283–1287.

31. AstraZeneca — Research-Based BioPharmaceutical Company. – URL: https://www.astrazeneca.com/ (дата обращения: 06.02.2024) – Текст: электронный.

32. AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning / S. Genheden, A. Thakkar, V. Chadimová [et al.] // Journal of Cheminformatics. – 2020. – Vol.12. №1. – P. 70.

33. The MIT License. – URL: https://opensource.org/license/mit/ (дата обращения: 07.02.2024) – Текст: электронный.

34. RDKit. – URL: https://www.rdkit.org/ (дата обращения: 07.02.2024) – Текст: электронный.

35. TensorFlow. – URL: https://www.tensorflow.org/ (дата обращения: 07.02.2024) – Текст: электронный.

36. ASKCOS. URL: https://askcos.mit.edu/login (дата обращения: 07.02.2024) – Текст: электронный.

37. AiZynthFinder GitHub. – URL: https://github.com/MolecularAI/aizynthfinder (дата обращения: 07.02.2024) – Текст: электронный.

38. ChEMBL Database. – URL: https://www.ebi.ac.uk/chembl/ (дата обращения: 08.02.2024) – Текст: электронный.

39. United States Patent and Trademark Office U. O. of P. Affairs (OPA) – URL: https://www.uspto.gov/ (дата обращения: 08.02.2024) – Текст: электронный.

40. ZINC20 — A Free Ultralarge-Scale Chemical Database for Ligand Discovery / J. J. Irwin, K. G. Tang, J. Young [et al.] // Journal of Chemical Information and Modeling. – 2020. – Vol.60. №12. – P. 6065–6073.

41. A robotic platform for flow synthesis of organic compounds informed by AI planning / C. W. Coley, D. A. Thomas III, J. A. M. Lummis [et al.] // Science. – 2019. – Vol.365. №6453. – P. 557–566.

42. Zhong, W. Retrosynthesis prediction using an end-to-end graph generative architecture for molecular graph editing / W. Zhong, Z. Yang, C. Y. C. Chen // Nature Communications. – 2023. – Vol.14. №1. – P. 3009.

43. Graph2Edits GitHub. – URL: https://github.com/Jamson-Zhong/Graph2Edits (дата обращения: 11.02.2024) – Текст: электронный.

44. Spaya. – URL: https://iktos.ai/ (дата обращения: 11.02.2024) – Текст: электронный.

45. IBM RoboRXN | Science | IBM Research. – URL: https://research.ibm.com/science/ibm-roborxn/ (дата обращения: 11.02.2024) – Текст: электронный.

46. rdchiral: Wrapper for RDKit's RunReactants to improve stereochemistry handling. – URL: https://pypi.org/project/rdchiral/ (дата обращения: 11.02.2024) – Текст: электронный.

47. Keras: API высокого уровня для TensorFlow. – URL: https://www.tensorflow.org/guide/keras (дата обращения: 12.02.2024) – Текст: электронный.

48. ONNX | Home. – URL: https://onnx.ai/ (дата обращения: 12.02.2024) – Текст: электронный.

49. Tree search — AiZynthFinder documentation. – URL: https://molecularai.github.io/aizynthfinder/relationships.html#tree-search (дата обращения: 12.02.2024) – Текст: электронный.

50. Analysis and post-processing — AiZynthFinder documentation – URL: https://molecularai.github.io/aizynthfinder/relationships.html#analysis-post-processing (дата обращения: 12.02.2024) – Текст: электронный.

51. Sequences — AiZynthFinder documentation. – URL: https://molecularai.github.io/aizynthfinder/sequences.html (дата обращения: 12.02.2024) – Текст: электронный.

52. Ertl, P. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions / P. Ertl, A.

Schuffenhauer // Journal of Cheminformatics. – 2009. – Vol.1. №1. – P. 8.

53. PubChem — the world’s largest collection of freely accessible chemical information. – URL: https://pubchem.ncbi.nlm.nih.gov/ (дата обращения: 01.04.2024) – Текст: электронный.

54. BLDpharm — Reliable research chemicals supplier. – URL: https://www.bldpharm.com/ (дата обращения: 01.04.2024) – Текст: электронный.

55. Angene Chemical. – URL: https://www.angenechemical.com/ (дата обращения: 01.04.2024) – Текст: электронный.

56. Pharmaceutical Chemicals | LEAPChem Chemical Supply. – URL: https://www.leapchem.com/ (дата обращения: 24.04.2024) – Текст: электронный.

57. Welcome to Click — Click Documentation (8.1.x). – URL: https://click.palletsprojects.com/en/8.1.x/ (дата обращения: 24.02.2024) – Текст: электронный.

58. Loguru GitHub. – URL: https://github.com/Delgan/loguru (дата обращения: 10.03.2024) – Текст: электронный.

59. Docker: Accelerated Container Application Development. – URL: https://www.docker.com/ (дата обращения: 10.03.2024) – Текст: электронный.

60. The most-comprehensive AI-powered DevSecOps platform | GitLab. – URL: https://about.gitlab.com/ (дата обращения: 10.03.2024) – Текст: электронный.

61. Kubernetes — Production-Grade Container Orchestration. – URL: https://kubernetes.io/ (дата обращения: 10.03.2024) – Текст: электронный.

62. Ubuntu 22.04.4 LTS (Jammy Jellyfish). – URL: https://releases.ubuntu.com/jammy/ (дата обращения: 13.03.2024) – Текст: электронный.

63. FastAPI. – URL: https://fastapi.tiangolo.com/ (дата обращения: 10.03.2024) – Текст: электронный.

64. TypeScript — JavaScript With Syntax For Types – URL: https://www.typescriptlang.org/ (дата обращения: 10.03.2024) – Текст:

электронный.

65. Datasets and their influence on the development of computer assisted synthesis planning tools in the pharmaceutical domain / A. Thakkar, T. Kogej, J. L. Reymond [et al.] // Chemical Science. – 2019. – Vol.11. №1. – P. 154–168.

66. The Open Reaction Database / S. M. Kearnes, M. R. Maser, M. Wleklinski // Journal of the American Chemical Society. – 2021. – Vol.143. №45. – P. 18820–18826.