Ретросинтетика: модель ретросинтеза от последовательности к последовательности

Инносент Джозеф
Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)
Innocentj36@gmail.com

Яна А. Бекенева
Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)
Yabekeneva@etu.ru

Онах Самуэль Адаколе
Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)
Sonakh@stud.etu.ru

Таофик Махфуз Мухаммад
Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)
Mtaofik@stud.etu.ru

Римвен Нансель Питер
Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)
rimvennansel77@gmail.com

Аннотация. **Оптимизация синтеза лекарств – постоянная проблема в фармацевтических исследованиях, обусловленная необходимостью ускорить открытие, сократить затраты и экспериментальные усилия, сохранив при этом безопасность и эффективность препарата. Это натолкнуло нас на идею создания фреймворка RetroSynthNet, который представляет собой оптимизированную модель "последовательность-последовательность", предназначенную для предсказания ретросинтетического маршрута с использованием набора данных USPTO. Наш подход использует механизм внимания и пользовательский токенизатор SMILES для эффективного преобразования входных молекулярных структур в жизнеспособные синтетические маршруты. Он был оценен по таким метрикам, как оценка BLEU, сходство по Танимото и химическая валидность, и продемонстрировал хорошую производительность при сохранении вычислительной эффективности. Данное исследование предоставляет практический инструмент для ускорения разработки лекарств и закладывает основу для будущих улучшений в области многоступенчатого ретросинтеза и расширенного химического прогнозирования.**

Ключевые слова: ретросинтез; глубокое обучение; последовательность к последовательности; лучевой поиск; SMILES

I. Введение

Оптимизация синтеза лекарств – одна из основных задач фармацевтических исследований, где необходимость ускорить открытие и снизить затраты пересекается с обеспечением безопасности и масштабируемости конечного продукта. Традиционные методы, основанные на ручном ретросинтетическом анализе и итеративных экспериментах, отнимают много времени и ресурсов, что часто приводит к затягиванию сроков разработки и неэффективности [1], [2]. Сложность современных терапевтических мишеней и обширное химическое пространство, в котором они обитают, делают крайне важным поиск более эффективных подходов. В одном из исследований подчеркивается, что традиционный синтез не справляется со сложными молекулами, что подталкивает к поиску решения с помощью искусственного интеллекта [2].

Появление искусственного интеллекта (ИИ), в частности моделей машинного обучения (МО), обученных на больших массивах данных о реакциях, таких как данные USPTO, изменило эту область. Прогнозирование ретросинтеза на основе ИИ использует нейронные сети для перевода молекулярных структур, часто представленных в виде SMILES-строк, в потенциальные реакционные прекурсоры [3], [4]. Эта технология обещает предсказывать синтетические маршруты, оценивать осуществимость молекул и исследовать новые химические вещества с непревзойденной эффективностью [5], [6].

Существующие методы прогнозирования ретросинтеза можно разделить на основанные на шаблонах, полушаблонные и бесшаблонные подходы [7], [8]. Модели на основе шаблонов, например, использующие библиотеки правил реакций, хорошо работают в структурированных условиях, но с трудом справляются с новыми реакциями, выходящими за пределы их предопределенной области [9], [10]; методы на основе полушаблонов разбивают проблему на предсказание синтонов и завершение реакций, включая химическую интуицию, но увеличивая сложность и зависимость от внешних баз данных. Методы без шаблонов, в первую очередь модели "последовательность-последовательность" (Seq2Seq), такие как молекулярный трансформатор, обеспечивают гибкость и масштабируемость, напрямую генерируя

реактивы из SMILES продуктов [11], [12]. Однако эти подходы часто сталкиваются с такими проблемами, как генерация химически недействительных структур и неспособность полностью использовать структурные и последовательные связи в молекулярных трансформациях, что приводит к неоптимальной точности и разнообразию предсказаний [11], [13], [14].

Несмотря на эти достижения, существует критически важный компромисс между вычислительной эффективностью и химической точностью при предсказании ретросентеза без шаблонов, который с трудом удается найти в существующих моделях. Во многих моделях Seq2Seq приоритет отдается предсказательной силе, а не практическому удобству, что приводит к созданию вычислительно-интенсивных систем, которые недостаточно используют богатую контекстуальную информацию в SMILES-представлениях [15]. Отсутствие специализированных механизмов токенизации и внимания мешает им улавливать тонкие зависимости между последовательностями продуктов и реактивов, что ограничивает их возможности в реальных рабочих процессах синтеза лекарств [2], [16].

В связи с этим возникает ключевой вопрос: Можем ли мы разработать оптимизированную модель без шаблонов, которая эффективно использует данные о молекулярных последовательностях для создания химически обоснованных и синтетически осуществимых маршрутов ретросинтеза? Для решения этой задачи мы представили RetroSynthNet, новую модель Seq2Seq, специально разработанную для предсказания ретросинтеза. Обученная на наборе данных USPTO-50K, RetroSynthNet включает в себя пользовательский токенизатор SMILES и механизм внимания, позволяющий превращать молекулы-производители в жизнеспособные синтетические прекурсоры с повышенной эффективностью и безопасностью. Уделяя особое внимание облегченной архитектуре и надежной химической проверке, RetroSynthNet достигает высокой производительности, измеряемой с помощью оценок BLEU, сходства Танимото и коэффициентов валидности, сохраняя при этом вычислительную простоту. Данное исследование не только предоставляет практический инструмент для ускорения синтеза лекарств, но и закладывает масштабируемую основу для будущих инноваций в области многоступенчатого ретросинтеза и фармацевтических открытий на основе ИИ.

II. Обзор литературы

A. Общие сведения о синтезе лекарств

Синтез лекарств является основой фармацевтических исследований, позволяя создавать терапевтические соединения с целенаправленным биологическим действием. Традиционные подходы, основанные на ретросинтетическом анализе, введенном Э. Дж. Кори в 1960-х годах, предполагают разложение сложных молекул на более простые предшественники. Хотя этот метод и был новаторским, он в значительной степени зависит от интуиции эксперта и сталкивается с ограничениями, связанными с комбинаторным взрывом потенциальных химических реакций и промежуточных продуктов. Исследование новых химических соединений часто ограничивается доступностью жизнеспособных синтетических путей, что приводит к потере возможностей для открытия новых лекарств [1]. Традиционный синтез часто сопряжен с высокими затратами, длительными сроками и значительным воздействием на окружающую среду, что обуславливает необходимость поиска более эффективных, основанных на данных альтернатив [17], [18].

B. Подходы к синтезу лекарств, основанные на искусственном интеллекте

Появление искусственного интеллекта (ИИ) изменило синтез лекарств, предоставив инструменты для прогнозирования результатов реакций, оценки осуществимости синтеза и оптимизации путей с поразительной точностью [5], [6]. Основные методы ИИ включают:

- Нейронные сети для прогнозирования реакций: Молекулярный трансформатор, модель на основе трансформатора, пересмотрел предсказание реакций, используя большие наборы данных для прогнозирования результатов с высокой точностью [8]. Нейронные сети на основе графовых трансформаторов (GTNN) расширяют эти возможности, представляя реакции в виде графов, что улучшает прогнозирование химической реактивности [20].
- Генеративные модели: Генеративные адверсарные сети (GAN) позволяют осуществлять молекулярный дизайн de novo, создавая новые соединения с заданными свойствами, но при этом они часто требуют обширной оптимизации [21].
- Планирование ретросинтеза: Такие инструменты, как Chematica, используют алгоритмы на основе графов для автоматизации ретросинтеза, используя эвристические правила для определения эффективных маршрутов. Модели обучения с применением усиления позволяют уточнить эти маршруты, уравновешивая такие факторы, как стоимость, выход и устойчивость [15].

C. Проблемы и недостатки в ретросинтезе, управляемом искусственным интеллектом

Несмотря на эти достижения, остаются серьезные проблемы. Методы, основанные на шаблонах, отлично работают в знакомых областях реакций, но с трудом справляются с новизной из-за своей зависимости от предопределенных правил [16], [22]. Полушаблонные подходы, которые разбивают ретросинтез на предсказание синтонов и достраивание реактивов, усложняются и зависят от внешних баз данных, что ограничивает их гибкость [23]. Бесшаблонные методы, такие как модели Seq2Seq, обеспечивают большую адаптивность, но часто дают химически недействительные структуры и не могут отразить сложные молекулярные связи [24]. Эти недостатки подчеркивают потребность в инновационных моделях, таких как RetroSynthNet, которые направлены на достижение баланса между эффективностью, точностью и химической достоверностью.

III. Методология

Оптимизация синтеза лекарственных препаратов требует инновационных подходов, позволяющих эффективно ориентироваться в обширном химическом пространстве и минимизировать зависимость от трудоемких экспериментов методом проб и ошибок. Для решения этих задач в данном исследовании используется система глубокого обучения, основанная на наборе данных USPTO Retrosynthesis Dataset из Therapeutic Data Commons (TDC). Эта система объединяет передовую предварительную обработку данных, нейросетевое моделирование и систематическую тонкую настройку для прогнозирования маршрутов ретросинтеза с повышенной химической обоснованностью и синтетической реализуемостью.

A. Сбор данных, предварительная обработка и создание признаков

Набор данных ретросинтезаUSPTO-50K был получен от TDC, который предоставил богатую коллекцию SMILES-строк химических реакций и связанных с ними типов реакций. Для обеспечения единообразия данных все SMILES-строки для реактивов (входов) и продуктов (выходов) были канонизированы с помощью RDKit, что позволило эффективно снизить вариативность представления. Набор данных подвергается тщательной очистке для удаления недействительных записей и дубликатов, а затем разделяется на обучающий, проверочный и тестовый наборы для эффективной и воспроизводимой разработки моделей.

Для того чтобы передать детали химических структур, был разработан собственный токенизатор SMILES. Этот токенизатор разбивает каждую строку SMILES на значимые лексемы, рассматривая выражения со скобками как единое целое, чтобы сохранить важную структурную информацию. Из этих лексем был составлен обширный словарь, включающий специальные символы (<SOS>, <EOS>, <PAD> и <UNK>), которые облегчают моделирование последовательностей. Полученные последовательности кодируются в виде индексов лексем и имеют единую длину, что обеспечивает совместимость с последующими архитектурами нейронных сетей.

B. Архитектура, вывод и оценка моделей глубокого обучения

Наш фреймворк основан на модели "последовательность-последовательность" (Seq2Seq) с механизмом внимания. Кодер, состоящий из слоя встраивания и многослойной долговременной памяти (LSTM), преобразует токенизированные входные SMILES в высокоразмерное пространство признаков, которое отражает сложные последовательные зависимости. Модуль внимания динамически взвешивает выходы кодера для создания векторов контекста, которые информируют декодер на каждом шаге предсказания. Декодер, также реализованный в виде многослойного LSTM, использует этот контекст в сочетании с ранее сгенерированными лексемами для итеративного построения целевого маршрута ретросинтеза. Обучение проводится с использованием принудительной обработки учителем и потери перекрестной энтропии, а производительность постоянно контролируется с помощью специального валидационного набора.

Cross-entropy Loss – это функция потерь, которая помогает оптимизировать точность генерации текста:

$$\mathcal{L} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \quad (1)$$

y_i = Истинный знак; $\hat{y}_i$=Прогнозируемая вероятность появления маркера.

Для вывода данных была использована стратегия жадного декодирования, а затем подход был усовершенствован с помощью декодирования с поиском пучка. Алгоритм поиска пучка исследует несколько последовательностей-кандидатов на каждом шаге декодирования, выбирая наиболее перспективные пути на основе суммарных логарифмических вероятностей. Постобработка с помощью RDKit гарантирует, что сгенерированные SMILES соответствуют действительным химическим структурам. Для оценки модели используется набор метрик, включая оценки Bilingual Evaluation Understudy (BLEU) на уровне предложений и корпусов, коэффициент достоверности химического состава, точность точного совпадения и сходство Танимото на основе отпечатков Моргана для оценки синтаксической точности и химической достоверности предсказаний.

Лучевой поиск, метод декодирования для генерации последовательностей, отслеживает топ-K наиболее вероятных последовательностей:

$$S_k = \arg \max_{v_1, \ldots, v_T}^{k} \sum_{t=1}^{T} \log P(y_t | y_{1:t-1}, X) \quad (2)$$

S = лучшая последовательность; y_t= жетон на временном интервале t; X = входная последовательность; K= (размер луча) – гиперпараметр, влияющий на разнообразие.

Сумма представляет собой общую log-вероятность последовательности

Гистограмма иллюстрирует распределение оценок BLEU на уровне предложений. Она показывает, что большинство сгенерированных шагов синтеза умеренно соответствуют эталонным шагам (0,5–0,7), в то время как некоторые отклонения указывают на области, требующие улучшения точности и согласованности. Числа на оси частот представляют собой количество предложений, попадающих в каждый диапазон оценок BLEU.

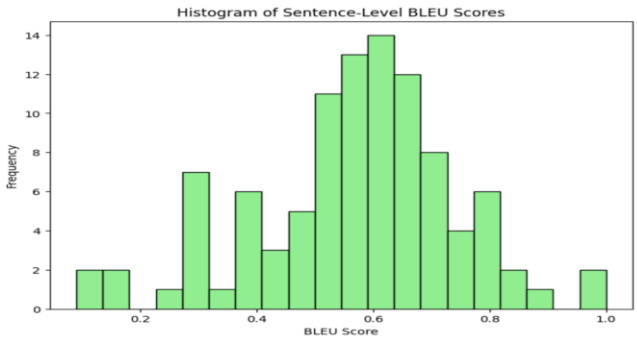

Рис. 1. Гистограмма BLEU

C. *Настройка гиперпараметров, тонкая настройка и будущие усовершенствования*

Понимая, что начальные конфигурации модели могут быть дополнительно оптимизированы, мы использовали систематический поиск по сетке для настройки критических гиперпараметров, таких как скорость обучения, отсев кодировщика и отсев декодировщика. Для каждой комбинации модель переобучалась в течение фиксированного числа эпох (10), и в качестве окончательной модели выбиралась конфигурация, дающая наименьшие потери при валидации. Эта фаза тонкой настройки не только улучшила общую производительность модели, но и повысила ее устойчивость, заложив основу для будущих усовершенствований. На левой гистограмме показаны баллы BLEU, свидетельствующие о том, что большинство сгенерированных шагов синтеза имеют умеренное выравнивание (0,5–0,7). На правой гистограмме представлены баллы сходства Танимото, показывающие, что предсказания молекулярной структуры в основном группируются около 0,3-0,6, что свидетельствует об умеренном химическом сходстве, но есть и высокоточные случаи (≥0,9).

Показатель Bilingual Evaluation Understudy (BLEU) оценивает сходство между этапами синтеза, сгенерированными ИИ, и эталонными этапами:

$$BLEU = BP.\exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \quad (3)$$

$BP = e^{(1-r/c)}$ – штраф за краткость для наказания за короткие предсказания; p_n = точность n-грамм; w_n = веса для различных n-грамм; r =контрольная длина; c =Длина кандидата.

Сходство Танимото измеряет структурное сходство между предсказанными и эталонными молекулами:

$$T(A,B) = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \quad (4)$$

$|A \cap B|$ =количество общих молекулярных признаков (например, субструктур, отпечатков пальцев); $|A|\ and\ |B|$ = общее количество признаков в каждой молекуле.

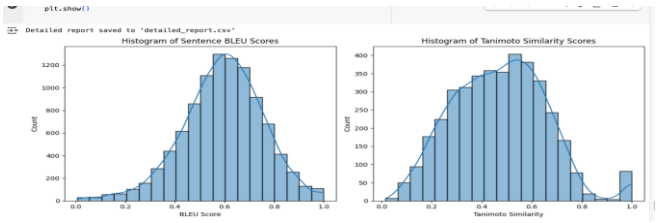

Рис. 2. Настроенные оценки BLEU

Этот систематический цикл от сбора данных до тонкой настройки модели напрямую устраняет ограничения традиционного планирования ретросинтеза. Автоматизируя предсказание маршрутов ретросинтеза и встраивая тщательную химическую проверку в процесс моделирования, наш подход значительно снижает экспериментальную нагрузку и расширяет возможности исследования возможных синтетических путей. В перспективе наша работа закладывает прочный фундамент для дальнейшего совершенствования, включая интеграцию обучения с усилением и специфических ограничений.

IV. Результаты, обсуждение и ограничения

Эксперименты показывают, что наш фреймворк глубокого обучения смог преобразовать набор данных ретросинтеза USPTO-50k в стабильную прогностическую модель. После тщательной предварительной обработки данных закодированные последовательности SMILES позволили эффективно обучить модель Seq2Seq с механизмом внимания. Используя декодирование с помощью лучевого поиска и тщательную химическую валидацию с помощью RDKit, модель достигла средней оценки BLEU на уровне предложений 0,73, оценки BLEU в корпусе данных 0,70 и химической валидности около 69 %. Кроме того, среднее сходство по Танимото между предсказанными и истинными молекулами составило 0,645, что отражает присущую ретросинтетическому предсказанию сложность. Детальный анализ ошибок, подкрепленный визуализацией оценок BLEU и распределений сходства по Танимото, диаграмма рассеяния визуализирует связь между выравниванием на основе текста и структурным сходством для сгенерированных шагов синтеза лекарств. Тенденция к росту показывает положительную корреляцию между оценками BLEU и сходством Танимото. Когда сгенерированный шаг синтеза имеет более высокий показатель BLEU, он также с большей вероятностью будет иметь более высокое молекулярное сходство.

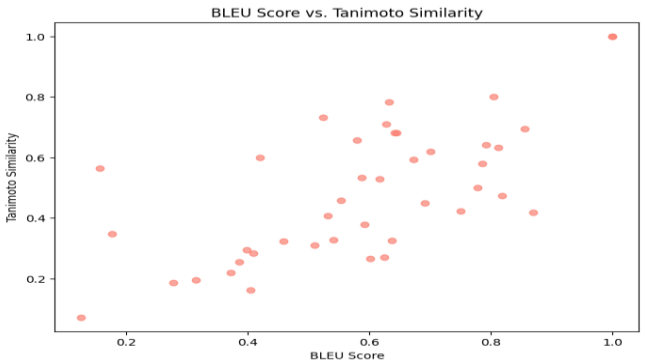

Рис. 3. Диаграмма рассеяния, BLEU и Танимото

Анализ ошибок показывает, что, хотя модель достигает умеренных показателей BLEU, некоторые сгенерированные молекулярные структуры не имеют химической обоснованности и недостающих значений сходства Танимото. Высокое текстовое выравнивание не всегда приводит к химически точным предсказаниям. Небольшой размер результатов был обусловлен значением поиска луча 5, что ограничивает разнообразие генерируемых кандидатов. Будущие усовершенствования должны включать химические ограничения и более широкие стратегии поиска для повышения достоверности, сохраняя при этом соответствие эталонным этапам синтеза.

Это подтверждает способность нашей системы улавливать основные химические и структурные особенности, создавая прочную основу для будущих усовершенствований, включающих стратегию обучения с усилением для дальнейшей оптимизации синтеза лекарств.

A. Обсуждение

Предсказание ретросинтетического маршрута остается важной задачей в вычислительном синтезе лекарств, требующей моделей, которые могут генерировать осуществимые строительные блоки молекул с эффективностью и надежностью. Традиционные методы, основанные на правилах, при всей их ценности часто зависят от экспертных знаний и предопределенных шаблонов реакций, что может ограничить их адаптивность к новым химическим пространствам [4]. Недавние достижения в области глубокого обучения позволили внедрить новые методологии. Некоторые исследователи предложили инновационное решение, структурировав ретросинтез как процесс молекулярной сборки, обеспечив повышенную интерпретируемость благодаря структурированной системе принятия решений. Однако у него есть определенные ограничения, в том числе трудности с предсказанием редких уходящих групп из-за зависимости от предварительно собранных данных, жесткий процесс принятия решений, который может не охватывать все типы реакций, и отсутствие детального предсказания условий реакции. Сильная зависимость от обучающих данных повышает риск переборщить с подгонкой, что сужает сферу его применения для новых реакций [4]. Другие исследователи также представили убедительный подход без шаблонов, включив выравнивание SMILES без надзора, что способствует улучшению предсказаний на основе последовательности. Несмотря на все преимущества, этот подход не позволяет учесть знания о домене, получить разнообразные результаты и справиться со случаями, когда структурное сходство продукта и реактанта слабое. Потенциальная утечка информации во время вывода может внести погрешности, что ограничивает его надежность в реальных приложениях [5].

RetroSynthNet предлагает оптимизированную схему Seq2Seq, в которой особое внимание уделяется эффективности и химическому правдоподобию. Наш подход позволяет избежать жестких структур принятия решений, а вместо этого дает возможность более адаптируемого прогнозирования путей ретросинтеза. Наша модель включает в себя этапы тщательной химической проверки с помощью RDKit, обеспечивая синтетически правдоподобные результаты, которые сохраняют осуществимость в широком диапазоне реакций. Ключевое преимущество нашей модели заключается в балансе между вычислительной простотой и глубиной прогноза, достигаемом с помощью легкого механизма, основанного на внимании. Однако наш подход не лишен ограничений. Текущая модель в первую очередь ориентирована на одностадийный ретросинтез, и хотя она хорошо справляется с предсказанием подходящих реактивов, в ней пока нет четкой информации об условиях реакции, которая необходима для синтеза в реальном мире. Наша модель не просто оптимизирует отдельные синтетические маршруты, она позволяет исследовать более широкое химическое пространство. Специальная токенизация и система глубокого обучения позволяют генерировать разнообразные ретросинтетические пути, создавая возможности для обнаружения новых соединений и реакционных путей, которые могли быть пропущены при использовании обычных методов. Выявление уникальных химических соединений может привести к прорыву в разработке терапии. Кроме того, как и в других моделях глубокого обучения, на их производительность влияет качество и разнообразие обучающих данных. В настоящее время персонализированная медицина и таргетная терапия приобретают все большее значение, поэтому способность эффективно исследовать и оптимизировать сложные химические пространства будет играть важную роль в разработке инновационных и экономически эффективных методов лечения. Наша система закладывает надежный фундамент для будущих инноваций, которые могут интегрировать передовые методы машинного обучения для планирования многостадийного синтеза, а совершенствование ее способности предлагать условия реакции может еще больше повысить ее применимость в открытии лекарств с помощью ИИ.

B. Ограничения

Наш метод показывает перспективность в предсказании маршрутов ретросинтеза и исследовании более широких химических пространств, но при этом остается несколько ограничений, которые указывают на направление будущей работы. Одна из проблем заключается в том, что примерно 69 % сгенерированных SMILES проходят проверку на химическую валидность, поэтому есть возможность улучшить то, насколько хорошо модель справляется со сложными деталями молекулярных структур. Мы рассматриваем это как возможность дальнейшего совершенствования путем интеграции более продвинутых архитектур. Мы намерены повысить химическую валидность, сохранив при этом преимущества эффективности.

Наши эксперименты были несколько затруднены из-за ограниченного доступа к высокопроизводительным вычислительным ресурсам. Это ограничило наши возможности по обучению и тонкой настройке более сложных моделей, что могло бы еще больше повысить производительность. Однако наш оптимизированный подход остается прочной основой для практического применения. При увеличении вычислительной мощности в будущих работах мы уверены, что модель может быть увеличена и расширена для работы с многоступенчатым ретросинтезом без ущерба для ее основных преимуществ.

V. Заключение

Данное исследование демонстрирует, что оптимизированная модель "последовательность-последовательность" может эффективно предсказывать маршруты ретросинтеза на основе базы данных USPTO, обеспечивая практический баланс между производительностью и вычислительной эффективностью. При этом остаются нерешенными такие задачи, как повышение химической обоснованности генерируемых SMILES и масштабирование модели на более современные архитектуры. Конкурентоспособные оценки BLEU и многообещающие меры сходства Танимото подчеркивают потенциал нашего подхода для ускорения синтеза лекарств. При дальнейшем совершенствовании и

расширении вычислительных ресурсов наш метод может проложить путь к более инновационной и экономически эффективной разработке лекарств.

СПИСОК ЛИТЕРАТУРЫ

[1] S. Талреджа и проф. д-р С. Тивари, "От одного к миллионам: Революция комбинаторной химии", J. Anal. Tech. Res., vol. 06, no. 02, 2024, doi: 10.26502/jatr.43.

[2] D. C. Blakemore et al., "Органический синтез предоставляет возможности для трансформации открытия лекарств", Nat. Chem., vol. 10, no. 4, pp. 383-394, Apr. 2018, doi: 10.1038/s41557-018-0021-z.

[3] R. Han, H. Yoon, G. Kim, H. Lee, and Y. Lee, 'Revolutionizing Medicinal Chemistry: Применение искусственного интеллекта (ИИ) в раннем открытии лекарств", Pharmaceuticals, vol. 16, no. 9, p. 1259, Sep. 2023, doi: 10.3390/ph16091259.

[4] Y. Wang et al., "Предсказание ретросинтеза с помощью интерпретируемого фреймворка глубокого обучения на основе задач молекулярной сборки", Nat. Commun., vol. 14, no. 1, p. 6155, Oct. 2023, doi: 10.1038/s41467-023-41698-5.

[5] K. Zeng et al., 'UAlign: Pushing the Limit of Template-free Retrosynthesis Prediction with Unsupervised SMILES Alignment', Apr. 19, 2024, arXiv: arXiv:2404.00044. doi: 10.48550/arXiv.2404.00044.

[6] B. Liu et al., 'Retrosynthetic Reaction Prediction Using Neural Sequence-to-Sequence Models', ACS Cent. Sci., vol. 3, no. 10, pp. 1103-1113, Oct. 2017, doi: 10.1021/acscentsci.7b00303.

[7] C. W. Coley et al., "Графо-конволюционная нейросетевая модель для прогнозирования химической реакционной способности", Chem. Sci., vol. 10, no. 2, pp. 370-377, 2019, doi: 10.1039/C8SC04228D.

[8] P. Шваллер и другие, "Молекулярный трансформатор: A Model for Uncertainty-Calibrated Chemical Reaction Prediction', ACS Cent. Sci., vol. 5, no. 9, pp. 1572-1583, Sep. 2019, doi: 10.1021/acscentsci.9b00576.

[9] 'AI-Enabled Retrosynthesis for Drug Development | Office of Innovation and Economic Development, The Ohio State University'. Accessed: Dec. 08, 2024. [Online]. Доступно: https://oied.osu.edu/technologies/ai-enabled-retrosynthesis-drug-development

[10] E. Garcia et al., "Cell-Line-Based Studies of Nanotechnology Drug-Delivery Systems", in Nanocarriers for Drug Delivery, Elsevier, 2019, pp. 375-393. doi: 10.1016/B978-0-12-814033-8.00012-6.

[11] ‘Overcoming Key Challenges in Drug Discovery’. Health and Medicine: Challenges for the Chemical Sciences in the 21st Century. Washington, D.C.: National Academies Press, 2004, p. 10889. doi: 10.17226/10889.

[12] D. Sun, W. Gao, H. Hu, and S. Zhou, 'Why 90% clinical drug development fails and how to improve it?', Acta Pharm. Sin. B, vol. 12, no. 7, pp. 3049-3062, Jul. 2022, doi: 10.1016/j.apsb.2022.02.002.

[13] Y. Jiang et al., 'Artificial Intelligence for Retrosynthesis Prediction', Engineering, vol. 25, pp. 32-50, Jun. 2023, doi: 10.1016/j.eng.2022.04.021.

[14] A.B. Садыбеков и B. Катрич, "Вычислительные подходы оптимизируют открытие лекарств", Nature, vol. 616, no. 7958, pp. 673-685, Apr. 2023, doi: 10.1038/s41586-023-05905-z.

[15] R. Bai, C. Zhang, L. Wang, C. Yao, J. Ge, and H. Duan, 'Transfer Learning: Making Retrosynthetic Predictions Based on a Small Chemical Reaction Dataset Scale to a New Level', Molecules, vol. 25, no. 10, p. 2357, May 2020, doi: 10.3390/molecules25102357.

[16] C.A. Shukla и A.A. Kulkarni, "Автоматизация многостадийного проточного синтеза: подход и проблемы интеграции химии, машин и логики", Beilstein J. Org. Chem., vol. 13, pp. 960-987, 2017, doi: 10.3762/bjoc.13.97.

[17] Более разумный способ ускорить открытие лекарств", MIT News | Massachusetts Institute of Technology. Accessed: Dec. 08, 2024. [Online]. Доступно: https://news.mit.edu/2024/smarter-way-streamline-drug-discovery-0617

[18] A.У. Рехман и др., "Роль искусственного интеллекта в революционном открытии лекарств", Fundam. Res., p. S266732582400205X, May 2024, doi: 10.1016/j.fmre.2024.04.021.

[19] D.F. Nippa et al., "Нейронная сеть с графовым трансформатором для прогнозирования химической реактивности", 16 мая 2023 г. doi: 10.26434/chemrxiv-2023-8hdmv.

[20] Y.J. Lee, H. Kahng, and S.B. Kim, 'Generative Adversarial Networks for De Novo Molecular Design', Mol. Inform., vol. 40, no. 10, p. 2100045, Oct. 2021, doi: 10.1002/minf.202100045.

[21] W. Zhong, Z. Yang, and C. Y.-C. Chen, 'Retrosynthesis prediction using an end-to-end graph generative architecture for molecular graph editing', Nat. Commun., vol. 14, no. 1, p. 3009, May 2023, doi: 10.1038/s41467-023-38851-5.

[22] I. A. Watson, J. Wang, and C. A. Nicolaou, 'A retrosynthetic analysis algorithm implementation', J. Cheminformatics, vol. 11, no. 1, p. 1, Dec. 2019, doi: 10.1186/s13321-018-0323-6.

[23] H. Qiao et al., 'Transformer-based multitask learning for reaction prediction under low-resource circumstances', RSC Adv., vol. 12, no. 49, pp. 32020-32026, 2022, doi: 10.1039/D2RA05349G.

[24] H. Duan, L. Wang, C. Zhang, L. Guo, and J. Li, 'Retrosynthesis with attention-based NMT model and chemical analysis of "wrong" predictions', RSC Adv., vol. 10, no. 3, pp. 1371-1378, 2020, doi: 10.1039/C9RA08535A.