УДК 004.896

РАЗРАБОТКА ВЕБ-СЕРВИСА ДЛЯ РЕТРОСИНТЕЗА НА ОСНОВЕ AI-ИНСТРУМЕНТА AIZINTHFINDER И УЛУЧШЕНИЕ КАЧЕСТВА ЕГО ПРЕДСКАЗАНИЙ

Голубев А.А.

ФГАОУ ВО «УрФУ имени первого Президента России Б.Н. Ельцина»,

Екатеринбург, Россия

argolubev97@yandex.ru

Аннотация. На основании AI-инструмента AiZynthFinder разработан удобный веб-сервис для предсказания и отображения интерактивных схем синтеза малых молекул. Также на двух наборах молекул (по 100 штук) проведено исследование факторов, влияющих на качество предсказаний.

Ключевые слова: искусственный интеллект, органическая химия, ретросинтетический анализ, AiZynthFinder.

DEVELOPMENT OF A WEB SERVICE FOR RETROSYNTHESIS BASED ON THE AI TOOL AIZINTHFINDER AND IMPROVEMENT OF ITS PREDICTIONS

Golubev A.A.

Ural Federal University, Ekaterinburg, Russia

Abstract. Based on the AI tool AiZynthFinder, a user-friendly web service has been developed for predicting and displaying interactive schemes for the synthesis of small molecules. Also, two sets of molecules (100 each) were used to study the factors affecting the quality of predictions.

Keywords: artificial intelligence, organic chemistry, retrosynthetic analysis, AiZynthFinder.

1. Введение

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 [1]. В последние годы ключевым способом автоматизации ретросинтетического анализа стало использование нейросетевых моделей. Одним из наиболее продвинутых AI-инструментов для многостадийного ретросинтетического планирования является AiZynthFinder [2].

Несмотря на все достоинства данного инструмента, он имеет ряд аспектов, требующих доработки. Во-первых, AiZynthFinder представлен только в виде python и cli интерфейсов, запуск которых требует от пользователя высокого уровня технической подготовки. Во-вторых, результаты его предсказаний «из коробки» часто оказываются неудовлетворительными. Однако, возможность работы с открытым кодом AiZynthFinder позволяет удобно интегрировать данный инструмент в собственные сервисы, а тонкая настройка параметров анализа помогает значительно улучшить результаты предсказаний.

2. Основная часть

AiZynthFinder – это open-source инструмент для построения схем синтеза малых молекул, основанный на применении трех нейросетей в рамках алгоритма поиска по дереву Монте-Карло [3]. На вход данному инструменту передается строчное представление целевой молекулы (SMILES), в качестве результатов предсказания выступают древовидные схемы синтеза. Целевая молекула считается успешно решенной алгоритмом в том случае, когда для неё найдена хотя бы одна схема, в которой все ветки дошли до соединений из стоков (сток – набор соединений, при достижении которых прекращается дальнейшее расширение ветки синтетической схемы).

В ходе данной работы на основании python интерфейса AiZynthFinder была написана собственная cli-утилита для запуска ретросинтетических задач на вычислительном кластере. Данная утилита производит инициализацию AiZynthFinder, запуск предсказания и обработку его результатов для отображения интерактивных схем синтеза в веб-интерфейсе вычислительного сервиса (рисунок 1). Кроме того, в рамках сервиса организовано хранение результатов анализа и удобный запуск предсказаний.

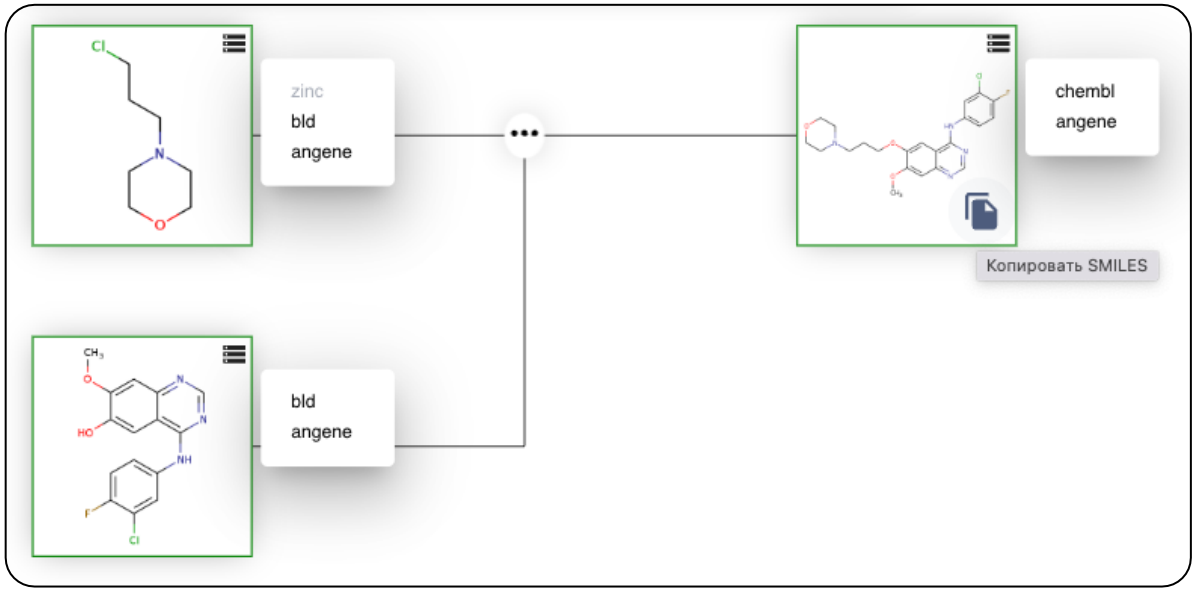

Рисунок 1 – Пример интерактивной схемы синтеза

Также было проведено исследование влияния параметров анализа на качество получаемых результатов. Инструмент тестировался в различных условиях на двух непересекающихся наборах молекул (по 100 штук). Было обнаружено, что наибольшее влияние на число успешно решенных алгоритмом молекул и, как следствие, качество итоговых предсказаний оказывает использование репрезентативных стоков (рисунок 2). Благодаря использованию стоков из различных источников и их комбинаций удалось увеличить долю решенных молекул почти в два раза (с 39–46 до 73–80%) по сравнению со стандартным стоком ZINC (zn). В качестве источников для создания стоков выступали открытые базы данных PubChem (pc), ChEMBL (cb); каталоги поставщиков LeapChem (lc), Angene Chemicals (ag), BLDPharm (bld); in-house данные (cs).

Наконец, на основании базы данных ORD [4] была обучена собственная ретросинтетическая модель, позволившая получать более предпочтительные схемы с точки зрения логики органического синтеза.

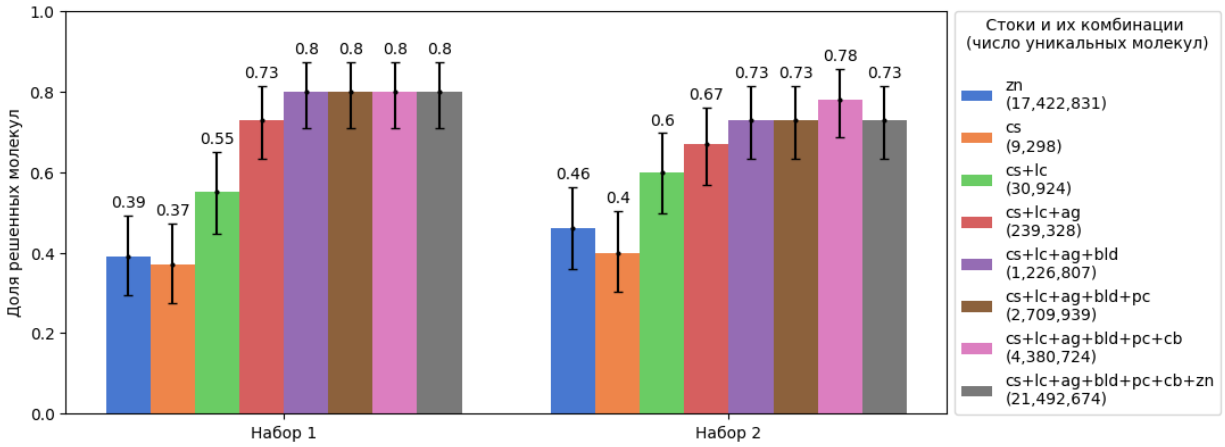

Рисунок 2 – Влияние стоков и их комбинаций на долю решенных молекул

3. Заключение

Таким образом, в рамках данной работы на основании AI-инструмента AiZynthFinder был разработан удобный веб-сервис для предсказания, отображения и хранения схем синтеза малых молекул. Благодаря проведенному тестированию и доработке инструмента удалось увеличить долю решаемых молекул с 39–46% (для стандартных условий) до 73–80%.

Библиографический список

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

2. Genheden, S. AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning // Journal of Cheminformatics. 2020. Т. 12. AiZynthFinder. № 1. – С. 70.

3. 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. – С. 604–610.

4. Kearnes, S. M. The Open Reaction Database // Journal of the American Chemical Society. 2021. Т. 143. № 45. – С. 18820–18826.