УДК 54.057

ПЛАНИРОВАНИЕ ОРГАНИЧЕСКОГО СИНТЕЗА С ПОМОЩЬЮ МЕТОДОВ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Орлова А.А. (Университет ИТМО), **Лавриненко А.К.** (Университет ИТМО),
Никитина К.В. (Университет ИТМО), **Сергеенко Е.А.** (Университет ИТМО)
Научный руководитель – профессор, д.х.н., Виноградов В.В.
(Университет ИТМО)

Введение. Ретросинтез является основополагающим методом для решения проблем, связанных с органическим синтезом. Цель ретросинтеза – найти наиболее оптимальный путь для получения целевого соединения путем его последовательного упрощения до более простых исходных веществ. В настоящее время для разработки последовательности реакций химики полагаются на свой опыт и знания – химическую интуицию. Основным недостатком такого подхода является предвзятость ученых, связанная с тенденцией использовать то, что уже было успешно в прошлом. Искусственный интеллект помогает объективно взглянуть на проблему ретросинтеза, основываясь на данных о миллионах ранее изученных органических реакций. Наиболее распространенными архитектурами нейронных сетей для решения задач ретросинтеза являются автоэнкодеры, трансформеры и графовые нейронные сети [1]. Однако, существующие алгоритмы не способны с достаточной точность подбирать условия для проведения реакций, а также не учитывают выход продукта, что является серьезным недостатком, поскольку зачастую даже незначительные изменения в условиях могут привести к резкому изменению состава продукта и его выхода. Целью данного исследования является разработка модели машинного обучения, способной подбирать условия реакций с учетом оптимизации выхода целевого соединения.

Основная часть. Для решения проблемы предлагается алгоритм, совмещающий в себе две модели: модель по предсказанию условий синтеза и модель по предсказанию выхода. Целевая молекула подается на вход модели по предсказанию условий, где генерируется набор параметров для проведения синтеза. На основании предложенных условий с помощью модели по предсказанию выхода определяется ожидаемый выход реакции. Предложенные условия ранжируются по величине предсказанного выхода продукта.

Для обучения моделей была собрана база данных органических реакций различных типов. Она содержит более 260 тысяч реакций, а также информацию о выходах продуктов и условиях: катализаторы, дополнительные реагенты, температура, давление, растворители. Для сбора данных из открытых источников был разработан автоматический парсер. Молекулы, участвующие в реакциях, были преобразованы в формат SMILES (Simplified Molecular-Input Line-Entry System), отражающий состав и структуру соединений. Для повышения точности предсказаний была проведена кластеризация реакций с помощью алгоритмов UMAP и HDBSCAN с целью определения типа трансформации. Для представления реакций был использован формат Differential Reaction Fingerprint [2], отображающий атомы и связи, участвующие в превращении. В результате кластеризации химических реакций были получены следующие классы: реакции окисления до карбоновых кислот, реакции восстановления кислот и спиртов, реакции образования простых эфиров, реакции присоединения азота и фосфора и др. Для органических молекул также были подобраны дескрипторы: молекулярная рефракция, поверхностное натяжение, ароматичность и количество галогенов составе структуры для растворителей; доступная для растворителя площадь поверхности и количество атомов металлов для катализаторов; дополнительные реагенты были классифицированы согласно их химической природе. На данный момент ведется разработка Message-Passing графовой нейронной сети [3] по предсказанию выхода на основе условий реакций.

Выводы. Собранная база данных является вариативной, поскольку содержит различные

типы реакций, а также широкий диапазон выходов реакций от близких к нулю до высоких. Разнообразие базы данных и полнота описания молекул способны обеспечить высокую точность модели. Разрабатываемый алгоритм может быть потенциально использован для более эффективного планирования синтеза в органических лабораториях.

Список использованных источников:

1. Dong J. et al. Deep learning in retrosynthesis planning: datasets, models and tools //Briefings in Bioinformatics. – 2022. – Т. 23. – №. 1. – С. bbab391.
2. Probst D., Schwaller P., Reymond J. L. Reaction classification and yield prediction using the differential reaction fingerprint DRFP //Digital discovery. – 2022. – Т. 1. – №. 2. – С. 91-97.
3. Gilmer J. et al. Message passing neural networks //Machine learning meets quantum physics. – 2020. – С. 199-214.