УДК 004.855.5

РАЗРАБОТКА СТРАТЕГИИ ПРОГНОЗИРОВАНИЯ ОРГАНИЧЕСКОГО СИНТЕЗА НА ОСНОВЕ МАШИННОГО ОБУЧЕНИЯ

Крюков А.Д. (Университет ИТМО)
Научный руководитель – аспирант, Лавриненко А.К.
(Университет ИТМО)

Введение. Продукты простых химических реакций, как правило, могут быть однозначно определены, однако для многих сложных органических реакций это является серьезной проблемой. Экспериментальный метод поиска продуктов реакции является времязатратным и требует помощи опытного химика. Существуют также квантово-химические способы, однако эти методы очень чувствительны к настройке параметров вычисления и требуют больших вычислительных ресурсов. Альтернативным подходом является применение методов машинного обучения, которые более масштабируемы и требуют меньше времени. Целью нашей работы является разработка алгоритма для предсказания продуктов органического синтеза, которые будут учитывать свойства реагентов, растворителей, а также катализаторов реакции.

Основная часть. Существующие на данный момент решения имеют ряд недостатков. Так, большинство моделей сфокусированы на определенных классах реакций, а в качестве дескрипторов используются только свойства [1] или символьное представление [2] молекул реагентов. Несмотря на то, что замена катализатора или растворителя может значительно влиять на результат реакции, эти параметры не учитываются в предсказательных моделях, что может вести к неточностям в определении продукта реакции.

Для разработки предсказательных алгоритмов нами был собран набор данных из патентной базы USPTO, состоящий из 149510 реакций. Датасет содержит информацию о реагентах, продуктах, катализаторах и растворителях органических реакций различных классов. Очистка и предобработка датасета производилась с помощью библиотек RDKit [3] и pandas и включала удаление дубликатов, проверку на достоверность с помощью базы данных SciFinder, и преобразование представления молекул в формат SMILES. На основе анализа данных, было выявлено, что катализатор присутствует у 45% всех реакций, а растворитель у 86%.

В своей работе для разработки алгоритма мы использовали свойства катализаторов и растворителей, характеризующие их молекулярную структуру, наличие функциональных групп, электронную структуру, поляризуемость, наличие доноров и акцепторов водорода, а также площадь полярной поверхности, полученные с помощью библиотеки RDKit. Для описания структуры веществ, участвующих в реакции, была использована токенизация и векторизация SMILES, полученная с помощью библиотеки word2vec. Для разработки алгоритма прогнозирования органического синтеза мы использовали архитектуры LSTM-RNN, Transformer-based with attention, BiLSTM. Для оценки точности модели использовался коэффициент Танимото, рассчитанный для предсказанного моделью и реального продукта реакции. Модель LSTM-RNN показала самую высокую эффективность предсказания из всех протестированных моделей. Ее Точность предсказаний составила 77%. Анализ важности дескрипторов производился с помощью методов SHAP и DeepLift. Наибольший вклад в предсказание продукта реакции оказывают дескрипторы, описывающие электронную структуру катализатора и молекулярную структуру растворителя.

Выводы. В ходе данной работы были исследованы различные подходы к применению искусственного интеллекта для прогнозирования продукта органических реакций. В качестве параметров для предсказания использовались структура реагентов, а также свойства катализаторов и растворителей. Модель LSTM-RNN с использованием модели word2vec для преобразования SMILES реагентов, растворителей и катализаторов показала высокую эффективность в прогнозировании продукта реакции. Точность предсказания оценивалась с

помощью коэффициента Танимото и составила 77%. Анализ важности дескрипторов показал, что электронная структура катализатора и молекулярная структура растворителя оказывают наибольшее влияние на предсказательную способность модели. Разрабатываемая нами нейронная сеть позволит прогнозировать продукты различных классов органических реакций, учитывая как реагенты, так и параметры синтеза.

Список использованных источников:

[1] Saini, V. Machine learning prediction of empirical polarity using SMILES encoding of organic solvents. Mol Divers (2022). https://doi.org/10.1007/s11030-022-10559-6

[2] Sheng Wang, Yuzhi Guo, Yuhong Wang, Hongmao Sun, and Junzhou Huang. 2019. SMILES-BERT: Large Scale Unsupervised Pre-Training for Molecular Property Prediction. In Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '19). Association for Computing Machinery, New York, NY, USA, 429–436. https://doi.org/10.1145/3307339.3342186

[3] Greg Landrum , RDKit Documentation Release 2019.09.1, May 13 2019 https://buildmedia.readthedocs.org/media/pdf/rdkit/latest/rdkit.pdf