**Поиск ошибок в хроматографических базах данных методами машинного и глубокого обучения**

***Хрисанфов М.Д.1,2***

*Аспирант, 3 год обучения*

*1Московский государственный университет имени М.В. Ломоносова,
химический факультет, Москва, Россия*

*2Институт физической химии и электрохимии им. А.Н. Фрумкина, Москва, Россия*

*E-mail:* *khrisanfovmike@gmail.com*

Базы данных времен удерживания METLIN SMRT и индексов удерживания NIST RI широко применяются при обучении предсказательных моделей и для обработки экспериментальных данных в высокоэффективной жидкостной и газовой хроматографии, соответственно. Для большей части веществ (более 80 %) в этих базах данных имеется только одно значение удерживания, поэтому для поиска ошибок невозможно использовать статистические методы. Проверить точность значений вручную также невозможно из-за большого объема данных. Ранее мы предложили подход для обнаружения потенциально ошибочных записей в базе данных NIST RI [1], а также адаптировали и применили его к METLIN SMRT [2].

Для нахождения ошибочных значений использовали 5 независимых предсказательных моделей. Предсказанные значения удерживания получали пятикратным повторением кросс-валидации с разбиением набора данных на 5 частей – 4 для обучения, 1 для предсказания. Таким образом для каждого соединения были предсказаны 5 значений удерживания. Экспериментальное значение удерживания рассматривали как потенциально ошибочное, если предсказания всех 5 моделей попадали в 5 % наименее точных одновременно в абсолютной и относительной шкалах. Мы обнаружили около 1500 (2 % от всех записей) потенциально ошибочных значений в базе данных METLIN SMRT [2] и 2093 (0.7 % от всех записей) - в базе данных NIST 17 RI [1].

Для METLIN SMRT были также проведены оценки воспроизводимости списка потенциально ошибочных значений путем обучения предсказательных моделей на разных тренировочных наборах. Для трех случайных разбиений пересечение между потенциально ошибочными записями, составило 88 %. Дальнейшая оценка эффективности предлагаемого подхода требует использования синтетических данных, что позволит проводить исследование в контролируемых условиях. Согласно предварительным результатам, точность (precision) нахождения ошибочных значений составляет более 90 %.

 *Данная работа выполнена при поддержке Министерства науки и высшего образования Российской Федерации в рамках госбюджетной темы № 124041900012-4.*

**Литература**

1. Khrisanfov M.D., Matyushin D.D., Samokhin A.S. A general procedure for finding potentially erroneous entries in the database of retention indices // Anal. Chim. Acta. 2024. Vol. 1297. P. 342375.

2. Khrisanfov M., Matyushin D., Samokhin A. Finding potentially erroneous entries in METLIN SMRT // J. Chromatogr. A. 2025. Vol. 1745. P. 465761.