

ПРОГНОЗИРОВАНИЕ ВРЕМЕНИ УДЕРЖИВАНИЯ В ОБРАЩЕННО-ФАЗОВОЙ ВЫСОКОЭФФЕКТИВНОЙ ЖИДКОСТНОЙ ХРОМАТОГРАФИИ С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

Ставрианиди А.Н.^{1,2}

¹*Химический факультет Московского государственного университета им. М.В.*

Ломоносова, 119991, г. Москва, Ленинские горы д. 1, стр. 3

²*Федеральное государственное бюджетное учреждение науки Институт физической химии и электрохимии им. А.Н. Фрумкина Российской академии наук, 119071, г. Москва,*

Ленинский проспект д. 31, корп. 4

stavrianidi.andrey@gmail.com

Одним из наиболее распространенных подходов к идентификации соединений в современном хроматографическом анализе является использование алгоритмов машинного обучения для оценки и прогнозирования параметров удерживания аналитов, включая коэффициент удерживания ($\log k$), коэффициент селективности (α) и время удерживания (t_R). Эти параметры играют ключевую роль в планировании эксперимента по разделению и идентификации соединений в сложных многокомпонентных смесях. Для обучения таких моделей используют наибольшую из доступных баз данных удерживания - METLIN SMRT. В данном исследовании используется расчет молекулярных дескрипторов в сочетании с процедурами выбора наиболее информативных признаков и последующим моделированием с использованием различных методов машинного обучения: многомерной линейной регрессии, метода опорных векторов, а также ансамблевых методов, в частности, случайного леса и его модификаций. Полученные модели предсказывают разделение фармацевтических веществ с использованием обращенно-фазовой ВЭЖХ. Метод случайного леса продемонстрировал наибольшую эффективность. Так, модель ChemoPy/RF правильно предсказала порядок элюирования структурно близких соединений и изомеров (компонентов четырех лекарственных форм) с точностью более 80%, используя в качестве входных параметров 3D молекулярные дескрипторы.

Для адаптации прогнозов времени удерживания к различным хроматографическим аналитическим условиям рекомендуется использовать методы глубокого обучения и трансферного обучения. Разработанная архитектура, основанная на одномерной сверточной нейронной сети, продемонстрировала высокую точность прогнозирования как на эталонном наборе данных METLIN SMRT (максимальная абсолютная ошибка около 35 с), так и при тестировании пяти независимых наборов данных. Альтернативой является разработка отдельных моделей, способных решать задачи специализированного группового прогнозирования удерживания и порядка выхода соединений с одинаковым структурным скелетом, например, стероидов и флавоноидов. Предложенный способ кодирования структуры гидрокси- и метоксизамещенных флавонов в виде строчного бинарного векторного представления позволил добиться высокой попарной точности (более 90%) предсказания их порядка элюирования. Интеграция методов машинного обучения в рутинный хроматографический анализ открывает возможности для повышения надежности идентификации компонентов сложных смесей, а также создает предпосылки для дальнейшего сокращения временных и материальных затрат путем ограничения в выборе возможных селективных систем и при оптимизации условий разделения для широкого спектра фармацевтических препаратов.

Исследование выполнено за счет гранта Российского научного фонда (проект № 22-13-00266-П), предоставленного Институту физической химии и электрохимии им. А. Н. Фрумкина Российской академии наук.