**Оценка склонности к переобучению в методах машинного обучения для расчёта электростатических свойств молекул**

***Зверев Д.В., Шульга Д.А., Шаймарданов А.Р., Палюлин В.А.***

*Аспирант, 1 г/о*

*Московский государственный университет имени М.В.Ломоносова,*

*химический факультет, Москва, Россия*

*E–mail: denzverev@gmail.com*

Модели машинного обучения демонстрируют высокое качество воспроизведения электростатических свойств молекул. Они обладают широким набором преимуществ, таких как гибкость, масштабируемость, скорость работы, а также высокая точность расчета сложных электростатических величин при оптимально подобранных параметрах. Данные методы являются перспективными для описания электростатических взаимодействий в молекулярном докинге и моделях молекулярной динамики [1].

Одним из основных недостатков любых моделей машинного обучения является склонность к переобучению – запоминание тренировочной выборки за счет высокой гибкости моделей, вместо выявления общих закономерностей в данных. Это приводит к ошибкам в прогнозировании свойств реальных соединений. Переобучение моделей может возникать за счет использования в качестве обучающей выборки неравномерных наборов структур. Большинство представленных в литературе моделей обучены на неравномерно распределенных по структурному разнообразию виртуальных библиотеках малых молекул, таких как ZINC, ChEMBL и QM9 [2].

В работе были отобраны модели машинного обучения, построенные на разных алгоритмах: случайный лес, градиентный бустинг, нейронные сети прямого распространения, графовые нейронные сети. Для анализа склонности к переобучению данных моделей была построена виртуальная библиотека структур малых органических молекул. Выборка была равномерно распределена и кластеризована по схожести молекулярных отпечатков с помощью пакета RDKit [3]. Для всех молекул был рассчитан электростатических потенциал (МЭП) с помощью квантово-химических методов в приближении RHF/6-31G\*. Отобранные модели были протестированы на созданной выборке, построено распределение по кластерам остаточной ошибки воспроизведения МЭП или частичных зарядов.

**Литература**

1. Fedik N. et al. Extending machine learning beyond interatomic potentials for predicting molecular properties // Nat. Rev. Chem. 2022. Vol. 6, № 9. P. 653–672.

2. Ramakrishnan R. et al. Quantum chemistry structures and properties of 134 kilo molecules // Sci. Data. 2014. Vol. 1, № 1. P. 1–7.

3. RDKit: Open-Source Cheminformatics Software, https://www.rdkit.org/ (accessed on September 1, 2023).