**Поиск оптимального набора эмбеддингов атомов для графовых нейронных сетей на основе химических языковых моделей**

***Остарков С.Н. 1,2, Долотказин М.C.1,2, Беспалов И.А. 1,2,3***

*Студент, 4 курс специалитета*

*1 Институт Органической химии имени Н.Д. Зелинского РАН, Москва, Россия*

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

*3Московский Государственный Технический Университет имени Н.Э. Баумана, Москва*

*E-mail: ostarkovstepan@gmail.com*

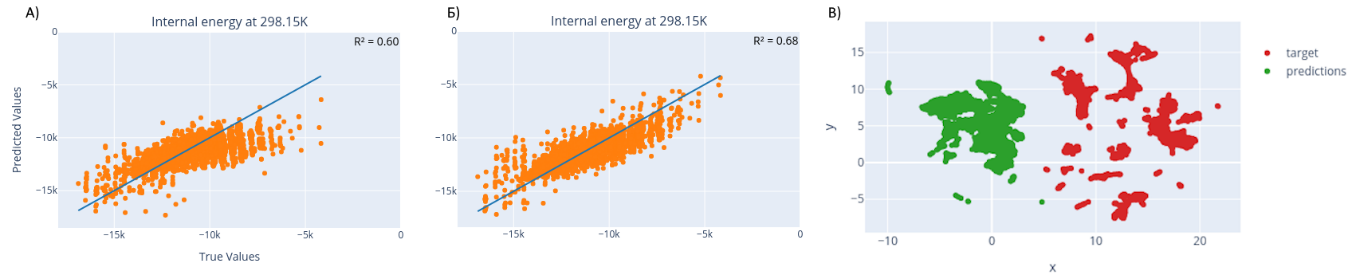
В последние годы методы машинного обучения получили значительное распространение в химических исследованиях. Графовые нейронные сети (ГНС) активно применяются для прогнозирования физико-химических свойств химических соединений и материалов. Однако проблема подбора оптимальных атомных эмбеддингов в таких моделях остаётся нерешённой и не имеет универсального подхода к реализации. Другие методы, основанные на языковых моделях, свободны от подобных ограничений. Тем не менее, подобные модели имеют меньшую обобщающую способность и не так эффективны для предсказания характеристик отдельных атомов.

Цель этой работы – создание ГНС, качество которой сопоставимо с крупными языковыми моделями с помощью метода дистилляции знаний. Полученные атомные эмбеддинги предполагается использовать в качестве универсальных для других моделей.

В качестве исходной языковой модели была выбрана ChemBERTa-2 [1], реализующая архитектуру трансформера, показавшая высокую эффективность в задачах языкового моделирования и многозадачной регрессии. Мы использовали ГНС на основе механизма графового внимания [2], имеющая в 4 раза меньше параметров, чем ChemBERTa-2. Для обучения использованы 150 000 структур из датасета модели ChemBERTa-2. Валидация эмбеддингов проводилась через их применение в качестве признаков для модели градиентного бустинга (ГБ), обученной для задачи многозадачной регрессии на датасете QM9 и сравнение с результатами обучения ГБ на основе эмбеддингов ChemBERTA-2.

Таблица 1. Значения R2 для различных задач регрессии на датасете QM9

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Модель | εHOMO | εLUMO | εHOMO-εLUMO | ZPVE | U298 | F298.15 | ΔatomF298.15 |
| **ChemBERTa-2** | 0.59 | 0.82 | 0.77 | 0.95 | 0.68 | 0.68 | 0.92 |
| **ГНС** | 0.46 | 0.58 | 0.52 | 0.72 | 0.60 | 0.59 | 0.66 |

Рис. 1. Сравнение качества эмбеддингов ChemBERTa-2 и ГНС  
**А** Предсказанные на основе эмбеддингов ГНС и целевые значения внутренней энергии; **Б** Предсказанные на основе эмбеддингов ChemBERTa-2и целевые значения внутренней энергии; **В** Сжатые с помощью UMAP эмбеддинги ChemBERTa-2 и ГНС

Анализ результатов показал, что качество эмбеддингов ГНС близко к целевым значениям на отдельных задачах регрессии. Для анализа распределения эмбеддингов был применён метод Uniform Manifold Approximation and Projection (UMAP) для снижения размерности. Показано, что эмбеддинги модели ГНС демонстрируют менее выраженную кластеризацию по сравнению с эмбеддингами языковой модели.

**Литература**

1. Ahmad W. et al. Chemberta-2: Towards chemical foundation models //arXiv preprint arXiv:2209.01712. – 2022.

2. Brody S., Alon U., Yahav E. How attentive are graph attention networks? //arXiv preprint arXiv:2105.14491. – 2021.