УДК 544.22

Активное обучение на основе графовых нейронных сетей с учетом синтонов для предсказания вероятности образования сокристаллов

Кадочникова М.С.[1]
Научный руководитель - аспирант Губина Н.В.[1]
[1]Университет ИТМО
kad.ms03@mail.ru

Введение

Сокристаллы позволяют изменять физико-химические свойства молекул, в том числе молекул лекарственных препаратов. Подбирать пары, которые образуют сокристалл экспериментально достаточно сложно и дорого. В данное время существующие методы на основе искусственного интеллекта не могут с достаточной точностью предсказывать вероятность сокристаллизации двух молекул [1]. В своей работе мы используем графовые нейросети и интегрируем информацию о супрамолекулярных синтонах, а именно о функциональных группах в молекулах, склонных к образованию межмолекулярных связей, приводящих к образованию сокристаллов. Это позволит уточнять прогнозы сокристаллизации молекул.

Основная часть

С помощью поисковой системы Google Scholar и ключевых слов “prediction cocrystals”, “cocrystals formation”, “formation probability” был произведен поиск научных статей, содержащих датасеты по образованию сокристаллических систем [1-4]. Сбор датасета проводился полуавтоматическим методом: программно, консолидация и обработка в Excel. Удалены дубликаты: SMILES переведены в каноническую форму. В некоторых работах было представлено стехиометрическое соотношение, а именно 144 пары AB и 32 пары A2B, которые также добавлены в датасет. Также была проведена фильтрация молекулярных структур из собранного датасета, в которую входило: органичность, количество атомов, молекулярная масса, количество циклов и количество тяжелых атомов. В качестве базовой модели написана простейшая графовая нейронная сеть. Далее была запущена исходная модель FragNet [5], где на вход подавались SMILES двух молекул через точку и результат, затем архитектура была изменена: на вход подаются SMILES двух молекул и результат. Также изменено программное преобразование датасета, которое включено в модель FragNet. На модифицированной модели FragNet получены высокие результаты: loss=0.1800, accuracy=0.9479, AUC=0.9718.

Выводы

Проведен обзор литературы. Данные из опубликованных статей были обобщены. Набор данных очищен/предварительно обработан путем сохранения только допустимых молекул и удаления дубликатов. Всего собрано 16307 молекулярных пар, из которых 14205 - образуют сокристаллы, а 2102 - не образуют. Визуализированы и проанализированы распределения набора данных. В качестве baseline написана простейшая графовая нейронная сеть, а также запущена модель Fragnet, в которой изменена архитектура для конкретной задачи. Результаты на модифицированной модели FragNet: loss=0.1800, accuracy=0.9479, AUC=0.9718.

Литература

1. Rebecca Birolo, Rıza Özçelik, Andrea Aramini, Roberto Gobetto, Michele R. Chierotti, Francesca Grisoni Deep Supramolecular Language Processing for Co-Crystal Prediction [Электронный ресурс]. – Режим доступа: https://onlinelibrary.wiley.com/doi/full/10.1002/anie.202507835 (Дата обращения 27.02.2026).
2. Yuanyuan Jiang, Zongwei Yang, Jiali Guo, Hongzhen Li, Yijing Liu, Yanzhi Guo, Menglong Li & Xuemei Pu Coupling complementary strategy to flexible graph neural network for quick discovery of coformer in diverse co-crystal materials [Электронный ресурс]. – Режим доступа: https://www.nature.com/articles/s41467-021-26226-7#citeas (Дата обращения 27.02.2026).
3. Dingyan Wang, Zeen Yang, Bingqing Zhu, Xuefeng Mei, and Xiaomin Luo Machine-Learning-Guided Cocrystal Prediction Based on Large Data Base [Электронный ресурс]. – Режим доступа: https://pubs.acs.org/doi/abs/10.1021/acs.cgd.0c00767 (Дата обращения 27.02.2026).
4. Lulu Zheng, Bin Zhu, Zengrui Wu, Mei Guo, Jinyao Chen, Minghuang Hong, Guixia Liu, Weihua Li, Guobin Ren, Yun Tang Pharmaceutical Cocrystal Discovery via 3D-SMINBR: A New Network Recommendation Tool Augmented by 3D Molecular Conformations [Электронный ресурс]. – Режим доступа: https://pubs.acs.org/doi/abs/10.1021/acs.jcim.3c00066 (Дата обращения 27.02.2026).
5. Gihan Panapitiya, Peiyuan Gao, C Mark Maupin, Emily G Saldanha FragNet: A Graph Neural Network for Molecular Property Prediction with Four Levels of Interpretability [Электронный ресурс]. – Режим доступа: https://arxiv.org/abs/2410.12156 (Дата обращения 27.02.2026).