УДК 544.22

COSYN: МУЛЬТИАГЕНТНАЯ СИСТЕМА ДЛЯ ДИЗАЙНА МЕТОДИК СИНТЕЗА СОКРИСТАЛЛОВ

Губина Н.В.[1], Баграмян А.С.[1], Кадочникова М.С.[1]
Научный руководитель – доктор технических наук Дмитренко А.В.[1]
[1]Университет ИТМО
gubina@pish.itmo.ru

Введение

Сокристаллы являются эффективным инструментом управления физико-химическими свойствами органических соединений, включая растворимость, стабильность и механические характеристики, что особенно востребовано в фармацевтической промышленности [1]. Несмотря на прогресс в применении искусственного интеллекта для дизайна сокристаллов, существующие подходы в основном фокусируются на предсказании вероятности совместной кристаллизации молекулярной пары [2] и предсказании свойств получаемого сокристалла [3]. Однако даже при выборе перспективной молекулярной пары исследователь сталкивается со следующим критическим этапом: получение сокристалла на практике часто требует систематического перебора широкого спектра экспериментальных условий, таких как растворитель, концентрация, соотношение компонентов, температура, время и способ воздействия. Данные о методиках синтеза разрозненны, представлены преимущественно в виде текстовых описаний в научных публикациях и не собраны в единый машиночитаемый источник. Это формирует потребность в масштабном извлечении и структурировании данных о синтезе и в разработке системы на основе искусственного интеллекта, способной рекомендовать или предсказывать параметры методик получения сокристаллов.

Основная часть

В работе построен пйплайн извлечения знаний о синтезе сокристаллов из научных публикаций и сформирован датасет параметров методик. На основе Кембриджской структурной базы данных [4] сформирован перечень статей, связанных с зарегистрированными сокристаллами, включающий 4 287 публикаций. По уникальным идентификаторам DOI выполнены автоматизированное получение материалов, преобразование файлов в текстовый формат и подготовка корпуса к последующей обработке.

Экстракция параметров синтеза сформулирована как иерархическая задача для больших языковых моделей. На первом этапе определяется наличие описания методики синтеза в тексте. На втором этапе классифицируется тип методики, включая твердотельное измельчение, суспензионные и растворные методы, а также другие подходы. На третьем этапе извлекаются параметры, специфичные для выбранного типа методики, в том числе растворитель и его количество, соотношение компонентов, температура и условия сушки. Для каждого этапа разработаны промпты и собраны валидационные датасеты объёмом 100 статей для первого этапа, 109 статей для второго этапа и 61 статья для третьего этапа. В сравнительном эксперименте протестированы модели GPT-4o-mini и Gemini 2.5 Flash. На всех этапах более высокие показатели качества продемонстрировала модель Gemini 2.5 Flash, которая использована для масштабной экстракции. В результате обработки 3 469 статей сформирован датасет, содержащий более 17 000 уникальных записей по 12 типам методик синтеза.

После формирования датасета выполнена постобработка, включающая удаление артефактов извлечения, унификацию категориальных значений, нормализацию единиц

измерения и согласование терминологии параметров между методиками. Дополнительно реализовано заполнение пропусков с учётом распределений параметров и контекста методик, что повышает пригодность данных для обучения и последующей валидации предсказательных моделей. Структура записей сохраняет связь параметров с молекулярной парой и источником публикации, что делает корпус пригодным для статистического анализа и обучения рекомендательных моделей.

Предсказание параметров методик синтеза выполнено с использованием больших языковых моделей и стратегии few shot, при которой ответ формируется на основе небольшого набора релевантных примеров. Для повышения точности применена мета-оптимизация промпта с разделением ролей критика и редактора, где критик выявляет неопределенности и ошибки, а редактор формирует уточненное предсказание. Следующими этапами работы являются расширенное тестирование предсказательной способности по ключевым параметрам и интеграция решения в мультиагентную систему, способную вести диалог, анализировать статьи и предлагать воспроизводимые протоколы синтеза.

Выводы

Разработан масштабируемый подход к сбору и структурированию знаний об условиях синтеза сокристаллов из литературы. Получен крупный корпус параметров методик, который закрывает дефицит единого источника данных и создает основу для построения моделей предсказания синтетических условий. Разрабатываемая система потенциально снижает объем экспериментального перебора, повышает воспроизводимость выбора методики и ускоряет получение сокристаллов для заданных молекулярных пар.

Литература

1. Bolla G., Sarma B., Nangia A. K. Crystal engineering of pharmaceutical cocrystals in thediscovery and development of improved drugs //Chemical Reviews. 2022. T. 122. No. 13. C.11514 -11603.
2. Jiang Y. et al. Coupling complementary strategy to flexible graph neural network for quickdiscovery of coformer in diverse co-crystal materials //Nature Communications. – 2021. – T.12. – No. 1. – C. 5950.
3. Gamidi R. K., Rasmuson Å. C. Analysis and artificial neural network prediction of meltingproperties and ideal mole fraction solubility of cocrystals //Crystal Growth & Design. –2020. – T. 20. – No. 9. – C. 5745-5759.
4. Cambridge Crystallographic Data Centre (CCDC). https://www.ccdc.cam.ac.uk