핵심 기술 연구 설계서 Organoid Source 설계

□ 연구 개요

○ 배경

- 오가노이드(Organoide) 기반 ADMET 데이터는 아직 체계적으로 큐레이션되어 있지 않아, 연구 및 모델링에 즉시 활용하기 어렵다는 한계가 존재한다.
- 관련 실험 정보들은 논문, 본문, 표, 부록, 캡션 등 비정형 형태로 분산되어 있고, 화합물 명칭/구조 표기, 단위, 실험 조건이 문헌마다 상이해 수집 및 정규화 비용이 크며 재현성이 낮다.
- 대규모 언어 모델(LLM)은 텍스트 및 이미지 혼합 문서에서의 정보 추출과 표준 스키마 정렬에 강점을 보인다. 본 연구는 LLM과 RAG, 표, 이미지 추출, 화학 구조 정규화를 결합하여 ADMET 예측 모델에 활용 가능한 데이터셋을 자동 구축하는 프레임워크를 설계 및 구현한다.

○ 관련 연구

- **TVAEs**[1]
  · NCBI 통합 검색, 접근 API로, PubMed 등에서 논문 메타데이터와 오픈 액세스 원문을 자동 수집하여 대규모 문헌 기반 데이터 구축에 활용 가능함
- **nanoMINER**[2]
  · LLM과 멀티모달 분석을 결합한 멀티 에이전트 시스템으로, YOLO(시각 추출)·GPT-4o(텍스트-시각자료 연계)·ReAct 조율을 통해 나노소재 문헌의 비정형 데이터를 자동 구조화하고, 화학식·결정 구조·표면 특성 등 핵심 파라미터를 정밀 추출함
- **ChemMiner**[3]
  · LLM 기반 멀티 에이전트 프레임워크로, 동지시어(Coreference) 해소를 핵심으로 하여 약어·기호·대용 표현을 단일 개체로 묶는 매핑 사전을 구축함. 이런 사전 기반 정규화를 통해 JSON 구조의 준수도(데이터 타입, 단위 일치, 값 범위, 형식 유효성)가 향상됨.
- **StructRAG**[4]
  · 질의 시점에 문서를 구조로 변환해 활용하는 RAG로, 사용자 질문을 의도 및 슬롯으로 분해하고 문서에서 얻은 구조화된 지식을 하이브리드 검색(Retrieval)으로 회수한 뒤, 사전 정의된 출력 규격에 맞춰 응답을 조립한다. 일관된 표/JSON 형식과 근거(출처)를 함께 제공함.

○ 기존 연구의 한계점 및 개선 사항

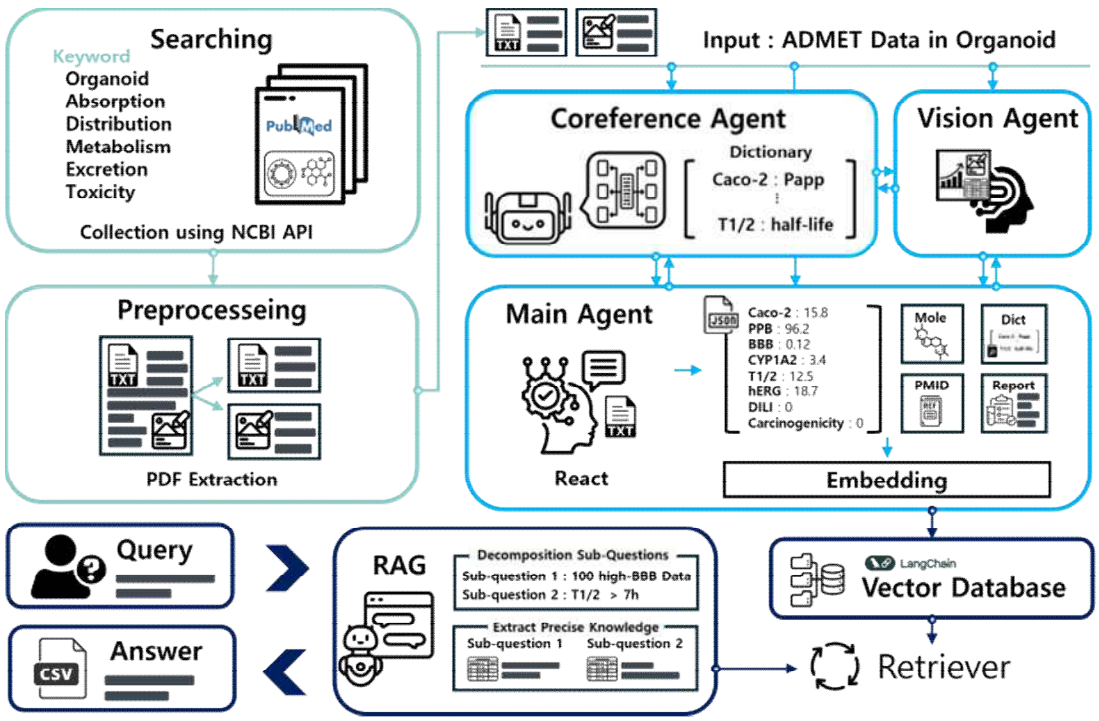
- 기존 연구들에서는 논문 곳곳에 흩어진 사진, 표 같은 비정형 데이터를 한 번에 모으고 정보를 정리하기 어려웠으며, 실험 결과들의 단위, 표기, 약어가 제각각이라 오류와 혼동이 많음.
- 이에 nanoMINER, ChemMiner 등은 실험 연구들에서 지식 추출을 보여주지만, 오가노이드 관련 ADMET에 맞춰진 기법이나 규칙이 부재하며, 지식 추출 단계에서만 멈추고, 사용자가 원하는 조건에 맞춰 재구성 해주는 단계가 결여됨.
- 특히, ADMET 예측 모델의 훈련 데이터로 사용하기 위해서는 문헌들에 나오는 수치를 자동으로 정규화 할 필요가 있으며, 자체적인 질의응답 시스템을 도입하여 원하는 데이터를 정확히 얻고 검증 할 수 있는 시스템이 필요함.

○ 연구 목표

- 이에 본 연구에서 수행하고자 하는 ADMET-RAG는 PubMed 논문을 자동으로 수집하고, 멀티 모달 지식 추출을 통해 본문, 표, 그림을 추출한 뒤 약어에 대한 사전 및 실험 데이터의 단위를 정규화하여 ADMET 전용 기법에 저장한다. 이후 구조화된 RAG를 이용하여 질의를 슬롯화 및 검증하여 훈련 데이터로 사용할 수 있는 CSV 파일과 파일에 대한 보충 설명 자료를 제공할 수 있는 프레임워크를 개발한다.

□ 주요 설계

○ ADMET-RAG 설계도



○ 수도 코드 (Pesudo code)

Input: User query for ADMET data
Output: ADMET dataset (CSV)

```
function ADMET RAG（Query):
    Fetch PDFs via Entrez
    for each PDF do
        Parse PDFs into texts, figures using layout analysis, OCR, and YOLO
        Extract FigureData from figures with the Vision agent
        Extract TextData form text with the Main agent
        Build a CorefDict from FigureData and TextData with the Coreference
        Generate JSON using FigureData, TextData, and CorefDict with the Main agent
        Chunk the JSON
        for each Chunk do
            Embed each chunk
            Upsert the embeddings into the vector database
        end for
    end for
    Decompose the query into sub-questions
    for each sub-question do
        Measure similarity and coreference to identify intent.
        Query the database with the RAG agent and generate schema-aligned answers
    end for
    Aggregate and review the answers
    return ADMET dataset (CSV)
```

□ 2단계 개발 일정

구분	내용	7	8	9	10	11	12
1	논문, 보충자료 수집 방법 구현	■	■				
2	논문 자료 전처리 자동화 구현		■	■			
3	Multi-Agent 오케스트라 구현			■	■	■	
4	ADMET-RAG 구현					■	■

□ 참고문헌

1. Database resources of the National Center for Biotechnology Information, Nucleic Acids Research, 38(Database issue), D5 – D16, 2010
2. Agent-based multimodal information extraction for nanomaterials, npj Computational Materials, 11, 194, 2025
3. ChemMiner: A Large Language Model Agent System for Chemical Literature Data Mining, arXiv preprint, 2024
4. StructRAG: Boosting knowledge intensive reasoning of LLMs via inference-time hybrid information structurization, ICLR, 2025