설계서 ADMET Ontology Structure

□ 연구 개요

○ 배경

- 대규모 언어모델(LLM) 기반 RAG를 신약개발 후보물질 탐색에 적용할 때, 방대한 ADMET·기전·부작용 정보를 모두 맥락에 반영하기에는 토큰 제한과 비정형 지식 분산으로 인해 정보 활용이 크게 제약된다. ADMET 온톨로지가 다양하게 구축되어 있음에도, RAG 과정이 복잡하고 비효율적이어서 후보물질의 핵심 속성을 일관된 구조로 연계·추론하기 위한 단순화·최적화 기술이 여전히 부족한 상황이다. 이에 따라 온톨로지 기반의 지식 구조를 경량화·정제하여 LLM이 필요한 정보를 효율적으로 검색·해석·추론할 수 있도록 하는 기술 개발이 필수적이다.

  · 신약개발 과정에서 ADMET(흡수·분포·대사·배설·독성) 특성은 후보물질의 성공 여부를 결정짓는 핵심 요소이지만, 실제로 실험 및 시뮬레이션으로 수집되는 ADMET 데이터는 분절되어 있고 서로 다른 데이터베이스 간 상관관계가 부족합니다. 대사 효소(CYP450 등)의 작용, 조직 내 약물 분포, 독성 반응 등은 복잡한 생물학적 인과관계망을 형성하나 이를 하나의 통합된 지식구조로 표현한 사례가 제한적입니다. [1] [2]

  · 온톨로지(ontology)는 이러한 복잡한 개념과 관계를 의미론적으로 표준화하고 논리적 추론을 가능하게 하는 구조적 틀을 제공하며, 생물의학 영역에서도 OBO Foundry 원칙을 기반으로 다수의 온톨로지 모델이 개발되어 왔습니다. 그런데 ADMET 영역에서는 약물·유전자·단백질·생체반응 등의 관계를 아우르는 ADMET 온톨로지 설계가 아직 초기 단계에 머물러 있어, 복수의 엔드포인트(ADMET 항목)와 다중 상관관계를 담을 수 있는 지식그래프 기반의 온톨로지 구축이 시급히 요구됩니다. [3] [4]

○ 관련 연구

- **Absorption (흡수)**

  · **ADMO (ADMET Domain Ontology)**: ADME 과정 중 흡수 단계를 구조화한 온톨로지. 위·소장·혈관 등 생리학적 기관과 수송체(Transporter) 개체를 포함

  · **ADMETlab 3.0**: ADMETlab의 HIA, Caco-2 permeability, P-gp substrate/inhibitor 등 주요 흡수 관련 변수를 OWL 구조로 변환한 사용자 정의 온톨로지

  · **pkCSM**: 약물 흡수 및 유전적 변이(transport, metabolism gene variant)에 대한 RDF

기반 데이터 제공

- Distribution (분포)
  · **DrugBank Ontology**: 약물의 체내 분포(Volume of Distribution, BBB permeability, Plasma Protein Binding) 관련 속성 정의
  · **ChEBI (Chemical Entities of Biological Interest)**: 화학적 엔티티와 생물학적 위치(Organ, Tissue) 간 관계를 표현하는 계층 구조 보유
  · **BRENDA Tissue Ontology (BTO)**: 인체 내 조직·기관 분류체계를 정의한 온톨로지로, 약물 분포 위치 표현에 활용

- Metabolism (대사)
  · **GO (Gene Ontology)**: Metabolic process, catalytic activity 등 대사 관련 생물학적 과정 정의
  · **PRO (Protein Ontology)**: CYP, GST, MAO 등 대사 효소 단백질 구조·기능·상호작용을 표준화
  · **Reactome / MetaCyc Ontology**: 생화학적 반응경로 및 대사산물 관계를 RDF/OWL 형식으로 제공
- Excretion (배설)
  · **ChEBI (chebi_full.owl)**: Metabolite, Excreted Product, Bile Acid, Urine 등 배설 관련 화합물 클래스 포함
  · **ADMO (ADMET Domain Ontology)**: enal clearance, Biliary excretion 등 생리학적 배설 경로를 시맨틱 구조로 표현
  · **Human Metabolome Database (HMDB)**: 체액(소변, 담즙, 혈액 등)에 존재하는 대사산물·배설산물 정보를 RDF 포맷으로 제공
- Toxicity (독성)
  · **OAE (Ontology of Adverse Events):** 약물 부작용, 독성 반응, 임상 이벤트 등을 계층적으로 정의한 공식 온톨로지
  · **TOX21 / EPA CompTox (RDF 변환 가능)**: 화합물의 세포독성·유전자독성 등 in vitro 실험결과를 RDF로 변환해 활용 가능
  · **ADMETlab Ontology Extension (admetlab.owl)**: SR-HSE, SR-MMP, SR-p53 등 독성 예측 모델 결과를 온톨로지 속성으로 표현
- Integrated Biomedical Knowledge Graphs
  · **DRKG (Drug Repurposing Knowledge Graph)**: DrugBank, PubChem, UniProt,

DOID 등 97개 DB 통합 RDF 그래프

· **Bio2RDF**: 생물학·화학·의학 데이터베이스를 RDF 기반으로 연결하는 글로벌 Linked Data 프로젝트

· **Hetionet**: 약물 – 유전자 – 질병 간 인과 관계를 통합한 생의학 지식 그래프

○ 기존 연구의 한계점 및 개선 사항

- 개별 속성 중심의 한계

· (기존 연구 한계) 기존 ADMET 관련 온톨로지(ADMO, OAE, GO, ChEBI 등)는 흡수·분포·대사·배설·독성 중 일부 속성만을 개별적으로 정의하고 있어, 전 과정을 통합적으로 표현할 수 있는 실무에 반영할 수 있는 프레임이 필요

· (개선 방향) ADMET 5대 속성을 상호 연계한 통합 스키마(Ontology Hub)를 정의하여, “Drug → ADMET_Property → Biological_Entity” 형태의 인과적 관계를 일관되게 표현하도록 설계

- 다중 온톨로지 활용의 비효율성

· (기존 연구 한계) 여러 온톨로지가 OWL/RDF 구조를 따르더라도 namespace·property·계층 구조가 서로 달라 RAG 구성 시 동일 개념을 탐색·매핑하기 위해 불필요한 정합성 검증과 중복 계산이 발생하며, 이는 후보물질 ADMET 지식 활용 효율을 크게 저하시킴

· (개선 방향) Bio2RDF·DRKG 등 통합 지식그래프의 스키마를 참고하여 prefix·URI·ObjectProperty를 표준화하고, 이를 기반으로 Cross-Ontology Alignment Framework를 도입함으로써 RAG 단계에서 다중 온톨로지 호출·매핑에 필요한 계산 비용을 최소화하여 활용 효율을 높임

- 데이터 기반 추론 한계

· (기존 연구 한계) 기존 ADMET 예측 모델(pkCSM, ADMETlab 등)은 통계·딥러닝 기반의 수치 예측에 집중되어 있으며, 예측 근거를 논리적으로 설명하거나 온톨로지와 연계하여 추론하는 구조가 부족함

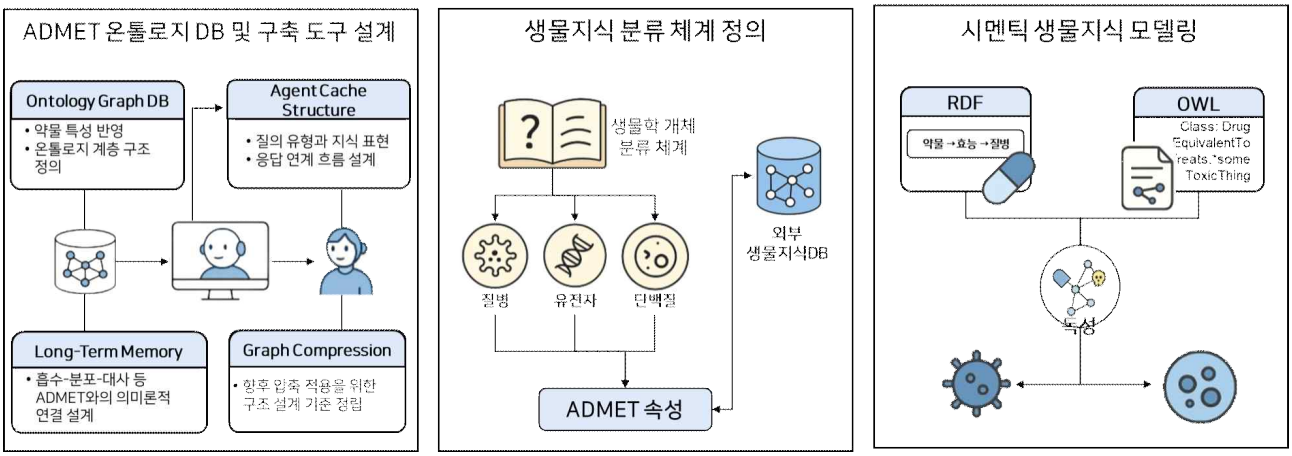
· (개선 방향) 예측 결과를 RDF 데이터 속성(data property)으로 변환하고, Reasoner 또는 LLM 기반 질의응답 시스템과 결합해 설명 가능한 추론형 지식 구조(Explainable ADMET-KG) 구축

○ 연구 목표

- 본 연구는 ADMET(흡수·분포·대사·배설·독성) 전 영역을 통합적으로 표현할 수 있는 온톨로지 기반 지식 구조를 설계하고, 이를 기반으로 약물 특성과 생물학적 개체(유전자·단백질·질병) 간의 의미론적 관계를 체계적으로 추론할 수 있는 지식그래프 데이터베이스(Knowledge Graph DB)의 구조를 정의하는 것을 목표

□ 주요 설계

○ ADMET 온톨로지 구축 방법 개요



○ 세부 설계 프로세스

- ADMET 속성별 온톨로지 구조 및 관계 정의

· ADMO, GO, OAE, ChEBI, ADMETlab.owl 등 핵심 온톨로지를 기반으로 Absorption – Distribution – Metabolism – Excretion – Toxicity 다섯 영역의 클래스(Class)·속성(Property) 계층을 분석하고, 각 속성 간 의미론적 관계(ObjectProperty: has_ADMET_property, metabolized_by, causes_toxic_effect 등)을 정의

· 약물(Drug) 개체와 생물학적 개체(Protein, Gene, Disease)를 연결하는 최소 RDF 삼중항 구조를 설계

- 상위 개념 통합 스키마(ADMET Hub Ontology) 설계

· 다섯 영역의 온톨로지를 통합 관리할 수 있는 상위 계층(Hub Ontology)을 정의

· 각 서브 온톨로지의 주요 클래스 간 관계를 통합하여 “Drug → ADMET_Property → Biological_Entity → Disease” 형태의 인과 구조를 표현

· Ontology 간 중복 용어를 제거하고, 공통 네임스페이스(prefix) 및 URI 체계를 표준화

- RDF 기반 지식그래프 데이터모델 정의

· 상위 스키마를 기반으로 RDF Triple 구조(Subject – Predicate – Object)를 정의하고, Triple Store(Fuseki, GraphDB 등)에 저장 가능한 데이터모델을 설계

· 데이터 속성(data property)으로 예측값·확률값(score) 등을 포함하여, AI 학습 및 추론 가능한 데이터 구조 설계

· RDF/XML 및 Turtle 형식(.ttl) 병행 설계로 확장성과 호환성을 확보

- User-Defined Language 나 SPARQL 질의 구조, 추론 규칙, 응답 설계

· ADMET 관계 탐색을 위한 표준 질의 구조를 정의

· OWL Reasoner(HermiT, Pellet 등)를 이용한 추론 규칙을 설계

○ 수도 코드 (Pesudo code)

Algorithm 1 Ontology Integration

1: Input: Ontologies O = {ADM O, OAE, GO, ChEBI, ADM ET lab.owl}
2: Output: Integrated RDF Graph G
3:**1. Load Ontologies**:
4:for all oi ∈O do
5:Import oi in RDF/OWL format
6:Extract core classes and object properties
7:end for
8:**2. Schema Mapping:**
9:Align namespaces and merge equivalent terms
10:Define unified schema H (ADMET Hub Ontology)
11: **3. Triple Generation:**
12:for all entity pair (Drug, Entity) do
13:Create triples using properties:
14:(Drug, has ADM ET property, P roperty)
15:(P roperty, related to, Entity)
16:end for
17:**4. Knowledge Graph Build:**
18:Store triples in RDF Store (e.g., Fuseki)
19:Apply Reasoner for ontology inference
20:**5. Query Setup:**
21:Define SPARQL templates for ADMET search
22:Export G as TTL/XML format
23:**return G**

○ 벤치마크

- 목적

· 본 벤치마크는 ADMET 온톨로지 기반 지식그래프 구축 과정에서 발생하는 대규모 RDF Triple 데이터 처리 성능, 그래프 구조 정제(Compression) 효율, 지식그래프 경량화에 따른 품질 변화를 정량적으로 검증

- 벤치마크 환경 (Benchmark Environment)

· 본 벤치마크는 Ubuntu 기반 로컬 환경에서 rootless Podman을 활용해 실제 서비스와 유사한 구조로 수행하였다. Micromamba 기반 Python 3.12.5 환경과 rdflib을 사용하여 대규모 TTL RDF 데이터를 스트리밍 방식으로 로드하고

Triple 단위 전처리를 진행하였다. 지식그래프 분석과 정제는 NetworkX 엔진을 중심으로 RDKit 등 파이썬 패키지를 결합해 ADMET 특성·화합물·유전자/단백질 관계를 유지한 채 그래프를 변환·통합하였다. ChEMBL 36.0 Activity RDF와 PC-Gene RDF를 활용해 약물 - 생체 반응 - 유전자 관계를 포함한 다중 RDF 소스를 처리하면서 CPU 기반 환경에서 메모리 및 처리 비용을 함께 계측하였다. 이러한 환경 구성은 RDF 변환 - 그래프 통합 - 그래프 정제를 일관된 조건에서 평가

항목	내용
OS	Ubuntu 기반 (Podman rootless 환경)
Storage	Local SSD (TTL 기반 RDF streaming 사용)
Python	3.12.5 (Micromamba)
RDF Parsing	rdflib
Graph Engine	NetworkX (구조 정제 및 압축 실험)
Toolkit	RDKit, tqdm 등

- 원본 데이터셋 구성 (Evaluation Dataset Overview)

· 원본 데이터셋 구성

구분	데이터셋	주요 내용	규모
RDF Dataset 1	ChEMBL 36.0 – Activity RDF	약물–표적 단백질–생체반응 관계, ADMET 연관 활성 값	약 1.9M ~ 3.5M triples
RDF Dataset 2	PC-Gene RDF	유전자 기능, 상호작용, 생물학적 경로 정보	약 1.8M ~ 3.4M triples

· 샘플링 전략 및 데이터 처리 방식

항목	내용
샘플링 방식	TTL 멀티라인 기반 Streaming Triple Sampling
도구	rdflib Streaming Parser

샘플링 수	각 RDF에서 6,000 triples, 총 12,000 triples
설계 원칙	구조 보존(연결 밀도, 시간적 순서, namespace 유지)

· 평가용 그래프 구성 결과

항목	내용
Proto-Graph 총 Edge 수	38,449
데이터 구성	ChEMBL(약물–표적–활성) + PC-Gene(유전자 상호작용) 통합
특징	원본 그래프 패턴을 유지한 축소형 ADMET-KG

- 평가지표 (Evaluation Metrics)

· 본 벤치마크에서는 ADMET 온톨로지 기반 지식그래프 구축 과정의 성능을 정량적으로 평가하기 위해 다음 네 가지 지표를 사용

항목	내용
RDF 스트리밍 속도 (Triple Sampling Performance)	대규모 TTL 형식 RDF 파일을 스트리밍 방식으로 읽어 들일 때, 단위 시간당 처리되는 Triple 수를 측정하여 RDF 파싱 및 샘플링 절차의 성능을 평가
그래프 병합 성능 (Graph Build / Merge Performance)	서로 다른 RDF 소스를 그래프로 변환한 뒤 통합하는 과정에서 생성되는 노드·엣지 수와 소요 시간을 계측하고, 데이터 증가에 따른 처리 부담을 분석함으로써 그래프 생성 및 병합 효율을 평가
그래프 정제·압축 성능 (Graph Compression Performance)	ENN, CLN, CLN-ADMET 등 단계별 그래프 정제 알고리즘 적용 전·후의 노드·엣지 수를 비교하여 압축률(감소율)을 산출하고, 불필요 관계 제거와 구조 경량화 효과를 정량적으로 측정
최종 그래프 효율성 (Final Graph Efficiency)	정제된 최종 그래프에 대해 메모리 사용량, 구조 복잡도, 탐색 비용 등을 종합적으로 분석하여, 향후 LLM-RAG 연계 시 질의 응답 성능 및 자원 활용 측면에서의 효율성과 확장성을 평가

- 벤치마크 결과 (Benchmark Results)

· RDF 스트리밍 및 Triple 샘플링 결과: ChEMBL Activity RDF와 PC-Gene RDF에 대해 각각 6,000개의 Triple을 스트리밍 방식으로 샘플링하였으며, 멀티라인 TTL 구조에서도 안정적인 파싱 성능을 보였으며, 이를 통해 총 12,000개의 정제된 Triple 세트를 생성하였고, 대규모 바이오 RDF 파일 처리 과정의 실용성을 확인

· 그래프 변환 및 통합 결과: 샘플링된 RDF 데이터를 NetworkX 기반 그래프로

변환한 뒤 통합하여 ADMET Proto-Graph를 구성하였으며, 통합 그래프는 총 38,449개의 엣지를 포함하고 서로 다른 RDF 소스를 결합하는 과정에서도 노드 및 엣지 구조가 안정적으로 유지되어 지식그래프 초기 구축에 적합한 성능을 확인

· 그래프 정제(Compression) 알고리즘을 통해 3단계 정제 알고리즘(ENN → CLN → CLN-ADMET)을 순차 적용한 결과, ENN 단계에서 19.73%, CLN 단계에서 57.29%, 최종 CLN-ADMET 단계에서 67.29%의 압축율을 달성

📊 Step 8 : 최종 정리 성능 요약

	방식	Nodes	Edges	Size(KB)	압축율(%)
0	Original	21040	38449	1384.16	-
1	ENN	17894	30862	1111.03	19.73%
2	CLN	7996	16423	591.23	57.29%
3	CLN-ADMET	6731	12578	452.81	67.29%

📝 Summary
- ENN : 정리율 19.73% → 기본 구조를 자연스럽게 정리하는 단계
- CLN : 정리율 57.29% → 문맥 단위로 구조를 다듬어 정보 흐름 개선
- CLN-ADMET : 압축율 67.29% → 최종적으로 활용하기 가장 적합한 균형 상태

- 결론

· 본 벤치마크를 통해 ADMET 온톨로지 기반 지식그래프 구축이 대규모 RDF 데이터에 대해서도 안정적으로 수행될 수 있고, 단계적 그래프 정제 알고리즘을 통해 최대 약 67% 수준의 구조 경량화가 가능함을 확인하였으며 이 과정에서 ADMET 특성 - 화합물 - 유전자/단백질 간 핵심 관계는 유지되면서 메모리 사용량과 탐색 비용이 감소하여, 향후 LLM-RAG 및 에이전트 기반 추론 시스템과 연계하기에 충분한 성능적 타당성을 확보

□ 1단계 1차년도 개발 일정

구분	내용	1	2	3	4	5	6	7	8	9	10	11	12
1	도메인별 온톨로지 사전조사												
2	ADMET 속성별 온톨로지 구조 사전조사												
3	온톨로지 간 매핑 규칙 설계												
4	상위 개념 스키마(Hub Ontology) 설계												
5	RDF 트리플 구조 사전조사												

□ 2단계 1차년도 개발 일정

구분	내용	1	2	3	4	5	6	7	8	9	10	11	12
1	온톨로지 도구를 통한 DB 생성 및 RDF 변환												
2	OWL 호환 GUI 편집 및 SPARQL 질의 기능 통합												
3	Plan-and-Execute 기반 캐시 조건 및 흐름 정의												
4	Agent가 참조 가능한 장기기억 구조 및 문맥 추출 로직 설계												
5	그래프 효율 향상 및 메모리 최적화 구조 설계												
6	RDF/OWL 구조 검증 및 질의 정확도 테스트												
7	Triple Store(Fuseki) 기반 질의 구조 구현												
8	OWL Reasoner(HermiT) 기반 관계 추론 검증												
9	Streamlit + pyvis 기반 지식그래프 UI 구축												
10	DB–Agent–LTM 모듈 통합 및 기능 검증												

□ 2단계 2차년도 개발 일정

구분	내용	1	2	3	4	5	6	7	8	9	10	11	12
1	실제 약물·질병·유전자 데이터 기반 온톨로지 DB 완성	■	■										
2	질의 유형별 온톨로지–LTM–캐시 연계 경로 구성			■	■								
3	응답 적중률, 기억 참조 정확도, 질의 일관성 측정					■	■						
4	Agent가 참조 가능한 장기기억 구조 및 문맥 추출 로직 설계							■	■				
5	AGI 기반 실증 질의 응답 프로토타입 구축									■	■		
6	사실성·정합성 검증 로직 및 위험 응답 차단 조건 정의											■	
7	Agent·LTM 간 데이터 호출 최적화												■

□ 2단계 3차년도 개발 일정

구분	내용	1	2	3	4	5	6	7	8	9	10	11	12
1	AGI 질의 흐름 연계 시 정확도·속도 기반 구조 보완	■	■										
2	다중 계층 구조 및 복합 질의 대응			■	■								
3	Vector DB·Key-Value Store 비교 및 성능 튜닝					■	■						
4	질의유형별 Agent 메모리 선택 경로 자동화							■	■				
5	위험 응답 차단·Self-refinement 구조 실증									■	■		
6	AGI 질의 시스템 통합 테스트 및 검증 모듈 완성											■	
7	최종 성능평가, 활용방안 및 기술이전 보고서 작성												■

□ 참고문헌

1. Li Fu et al., ADMETlab 3.0: an updated comprehensive online ADMET prediction platform, Nucleic Acids Research, vol.52, W422–W431, 2024
2. Gintautas Kamuntavičius et al., Benchmarking ML in ADMET predictions: the practical impact of feature representations in ligand-based models, Journal of Cheminformatics, 17:108, 2025
3. Mary Shimoyama et al., Multiple Ontologies for Integrating Complex Phenotype Datasets, Nature Precedings, 2009
4. Vincent Henry et al., Converting Alzheimer's disease map into a heavyweight ontology: a formal network to integrate data, arXiv preprint, 2018