핵심 기술 연구 설계서 **Knowledge COT 설계**

□ 연구 개요

○ 배경

- ADMET 예측을 위한 AGI 개발을 위해서는 학습 데이터 수집 뿐만 아니라, LLM 구축을 위한 Chain of Thought(CoT) 기반의 knowledge 학습데이터가 필요함
- 특히 ADMET 예측에 대한 단계적인 사고를 위해서는 예측에 핵심적인 생물학적인 매커니즘 설명과 다양한 사전 정보를 함께 학습해야하는 등 멀티모달 수준의 LLM을 지원할 수 있는 데이터 구축이 필요함
- 1차년도 본 연구에서는 아래와 같은 목표를 통해 연구를 수행함
  ① Chain of Thought(CoT) 기반 세포독성 예측 LLM 베이스라인 구축
  ② Tox21로부터 (화합물, 목표 단백질, 독성 레이블) 튜플로 이루어진 원시데이터 구축
  ③ 목표 단백질의 GO Term, KEGG pathway 등 생물학적 정보를 이용해 LLM 모델의 추론 학습을 위한 원천데이터 구축

○ 기존 연구의 한계점 및 개선 사항

- 화합물의 ADME/T 실험 결과는 신약개발 단계에서 화합물의 안정성 평가 등에 반드시 필요한 정보로 매우 높은 가치를 지니지만 데이터 수집을 위한 실험단계에서 큰 비용이 필요하여 Clinical Trial 단계까지 진행된 약물 평가를 위한 실험이 아니라면 보통 체계적으로 수행하지 않음
- 따라서 정보량 자체가 현저히 적을 뿐만 아니라 실험된 결과도 공개되지 않는 경우가 많고, 특히 공개되더라도 데이터베이스화 되지 않고 파편화되어있어 대용량의 학습 데이터가 필요한 딥러닝 기반의 연구가 쉽게 접근하기 어렵다는 한계가 뚜렷함
- 이중 Tox21 데이터베이스는 거의 유일하게 ADME/T 정보들 중 파편화되지 않고 정형화되어 딥러닝 학습에 쉽게 이용 가능함에도 여전히 QSAR등의 기초 머신러닝 기반의 연구들 위주로만 사용되고 있음

○ 연구 목표

- Tox21 데이터를 LLM을 이용한 CoT 기반 독성 예측 모델의 학습 데이터로 활용
- Tox21에서 제공되는 원시데이터 확보 및 데이터 전처리
- Knowledge CoT 학습을 위해 GO Term와 KEGG Pathway 정보를 활용한 원천데이터로 가공

□ 주요 설계

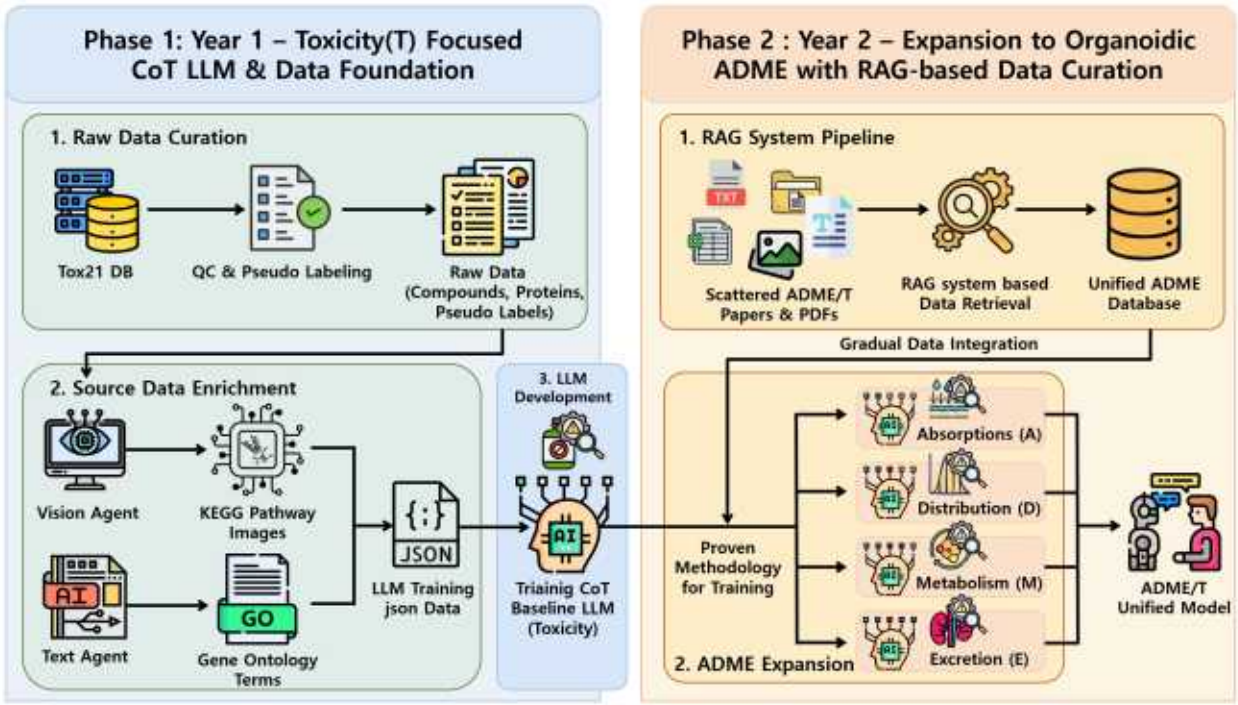
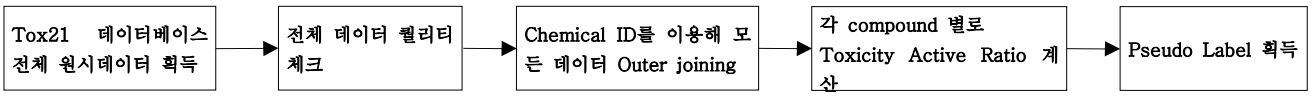


그림 1 제안하는 Knowledge CoT 설계를 위한 전체적인 설계도 (1, 2단계로 나누어 구성)

○ 원시데이터 큐레이션



1. Tox21 데이터베이스 원시데이터 수집

- Tox21 공개 데이터베이스에서 각 assay별 원시 csv 파일을 모두 다운로드함. 각 파일은 SAMPLE_ID, ASSAY_OUTCOME, CHANNEL_OUTCOME, PUBCHEM_CID, SMILES 등 주요 실험 결과와 화합물 정보를 포함

2. 전체 데이터 품질 관리(Quality Control)

- PURITY_RATING 컬럼을 이용하여 {A, Ac, B, Bc, C, Cc} 등급에 해당하지 않는 시료는 모두 제외하여 저품질 샘플을 제거
- 하나의 ASSAY에서 동일한 PUBCHEM_CID가 여러 번 측정된 경우, ASSAY_OUTCOME 또는 CHANNEL_OUTCOME에 'agonist'/'antagonist'가 포함된 row를 우선적으로 남기고 나머지 중복을 제거하여 화합물별 대표 실험값을 정제
- ASSAY_OUTCOME과 CHANNEL_OUTCOME을 기반으로 실험 결과를 정량화하기 위해 ToxicityScore 컬럼을 생성하였고, Tox21에서 권장하는 ASSAY_OUTCOME 우선 규칙을 적용하여 각각 0~2 또는 0~0.66 범위의 일관된 독성 점수로 변환

3. Chemical ID(PUBCHEM_CID)를 이용한 데이터 통합 (Outer Join)

- 각 ASSAY별로 정제된 csv에서 PUBCHEM_CID, SAMPLE_NAME, SMILES, ToxicityScore만 추출
- PUBCHEM_CID를 기준으로 모든 ASSAY결과를 outer join하여 하나의 compound - assay matrix (PUBCHEM_CID × assay)를 구성

- 이 과정에서 특정 화합물이 어떤 ASSAY에서 관측되지 않은 경우 해당 셀은 결측값(Null)으로 유지

4. Compound별 Toxicity Active Ratio / Score 계산

- 각 row(하나의 compound)에 대해 모든 assay의 ToxicityScore 값과 결측 비율을 이용해 다음과 같은 pseudo score를 계산:

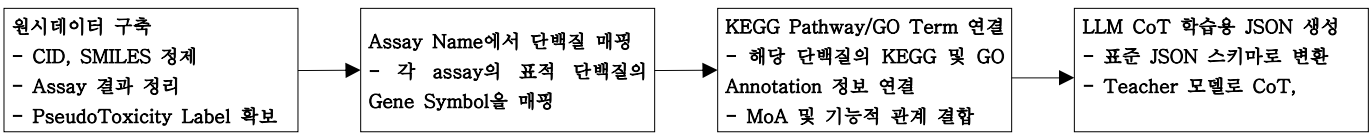
$$\text{PseudoScore} = \frac{\sum \text{ToxicityScore}}{\text{비결측 assay 수}} \times (1 - 0.5 \times \text{Null 비율})$$

- 위 식은 (1) 관측된 실험들에서의 평균적인 활성 정도와 (2) 전체 assay 중 실제로 측정된 비율을 동시에 반영하여, 데이터 sparsity에 대한 패널티를 부여하도록 설계

5. 최종 Pseudo Label 산출

- 계산된 PseudoScore에 대해 사전 정의한 임계값(0.5)을 기준으로 이진 라벨을 부여
- PseudoScore ≥ 임계값인 화합물은 독성 가능성이 높다고 판단하여 PseudoLabel = 1 (toxic)으로, 그 미만인 경우 PseudoLabel = 0 (non-toxic)으로 정의
- 이렇게 생성된 PseudoLabel은 후속 LLM 기반 독성 예측 모델의 학습용 레이블로 사용

○ 원천데이터 구축



1. Tox21 매트릭스 기반 원시데이터 정리

2. Assay-단백질 매핑 및 KEGG / GO 주석 연결

- (1) AssayName → 표적 단백질 매핑
- (2) KEGG Pathway 정보 연동 + 'Pathway 이미지'해석 활용
- (3) GO Biological Process / Molecular Function 연동

3. Biology Context 및 Local Causal Graph 생성

- 2단계에서 수집한 KEGG image-derived 정보 + GO term을 종합하여 짧은 contextual text를 생성
- 화합물의 작용 방향성(activation/inhibition) 단백질이 pathway에서 갖는 기능을 기반으로 Local Causal Graph Skeleton 생성하여 다음과 같은 구조로 JSON에 기록:

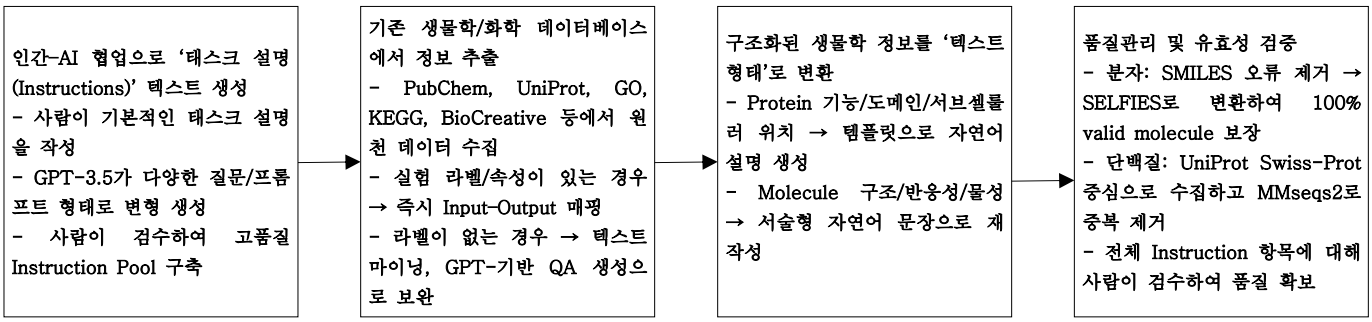
```
"causal_graph": {
"nodes": ["compound", "TP53", "BAX", "apoptosis", "toxicity"],
"edges": [
  {"source": "compound", "target": "TP53", "type": "inhibition"},
  {"source": "TP53", "target": "BAX", "type": "activation"},
  {"source": "BAX", "target": "apoptosis", "type": "activation"},
  {"source": "apoptosis", "target": "toxicity", "type": "positive_association"}
]
}
```

4. LLM CoT 학습용 JSON 변환

- Compound, Protein, KEGG/GO context, Causal Graph, Tox21 signal을 하나의 JSON으로 구성
- 이후 teacher LLM(GPT-4/5)을 이용하여 cot_steps, toxicity_label, short_rationale를 생성해 fine-tuning 데이터를 완성

○ LLM 사전학습용 데이터 구축

- Mol-Instructions에서 제공하는 LLM 사전학습을 위한 ① 분자, ② 단백질, ③ 생물학 텍스트 데이터세트로 도합 204만개의 텍스트 데이터를 사용
- Mol-Instructions 논문의 데이터 획득 절차는 다음과 같은 4단계로 구성



□ 2단계 개발 일정

구분	내용	7	8	9	10	11	12
1	논문, 보충자료 수집 방법 구현	■	■				
2	원시데이터 큐레이션		■	■			
3	원천데이터 구축			■	■	■	
4	LLM 사전학습용 데이터 구축					■	■