

□ 연구 개요

○ 배경

- 신약 후보 발굴을 위해 새로운 화학적 구조와 약리학적 특성을 가진 화합물의 지속적 탐색이 필요
- 기존 화합물의 성능 한계를 극복하고, 난치성 질환 치료 및 새로운 치료제 개발을 위해 새로운 화학 구조발굴이 필수적임
- 그러나 기존의 독성 평가 방식은 실험 기반으로 진행되기 때문에 시간·비용이 많이 들고, 초기 스크리닝 단계에서 대규모 후보군을 평가하는 데 한계를 지님. 또한 고전적인 머신러닝 모델이나 예측기반 QSAR 모델은 예측값만 제공할 뿐, 왜 특정 분자 구조가 독성을 유발하는지에 대한 설명(설명가능성, XAI)이 부족하여 규제 대응과 신뢰성 확보가 어려운 상황임.
- 이러한 배경에서 화학 분자 구조 정보를 기반으로 독성 여부를 예측하고, 그 판단 근거를 자연어로 설명할 수 있는 AI 기반 독성 추론 모델의 필요성이 커지고 있음.
- 특히 최근의 LLM 기술 발전은 단순한 예측을 넘어 추론(reasoning)의 수행을 가능하게 함. 그 중에서도 Chain-of-Thought(CoT) 형식은 모델이 판단 절차를 단계별로 서술하도록 유도하여, 독성의 원인·기전·구조적 특징을 보다 투명하게 드러내는 데 효과적인 기술임.
- 이러한 이유로, 분자 구조 기반의 독성 추론 데이터를 활용한 LLM 모델은 단순한 예측 도구를 넘어 설명 가능한 형태로 발전가능성을 지님.

○ 관련 연구

- CoTox (2025): CoTox는 분자 독성 예측에 LLM 기반 Chain-of-Thought(CoT) reasoning을 도입한 모델로, 화학 구조 정보와 생물학적 지식을 함께 활용하여 독성 여부와 단계적 추론 과정을 동시에 생성하는 모델임.
- An Explainable Supervised Machine Learning Model for Predicting Respiratory Toxicity of Chemicals Using Optimal Molecular Descriptors (2022) : 이 모델은 분자 구조에서 추출한 descriptor 기반 특징을 사용해 호흡기 독성 여부를 예측하는 supervised ML 모델임. 또한 SHAP과 같은 XAI 기법을 활용해 독성 예측에 기여한 중요 특징(feature importance)을 제시함으로써 설명 가능성을 제공함.

○ 기존 연구의 한계점 및 개선 사항

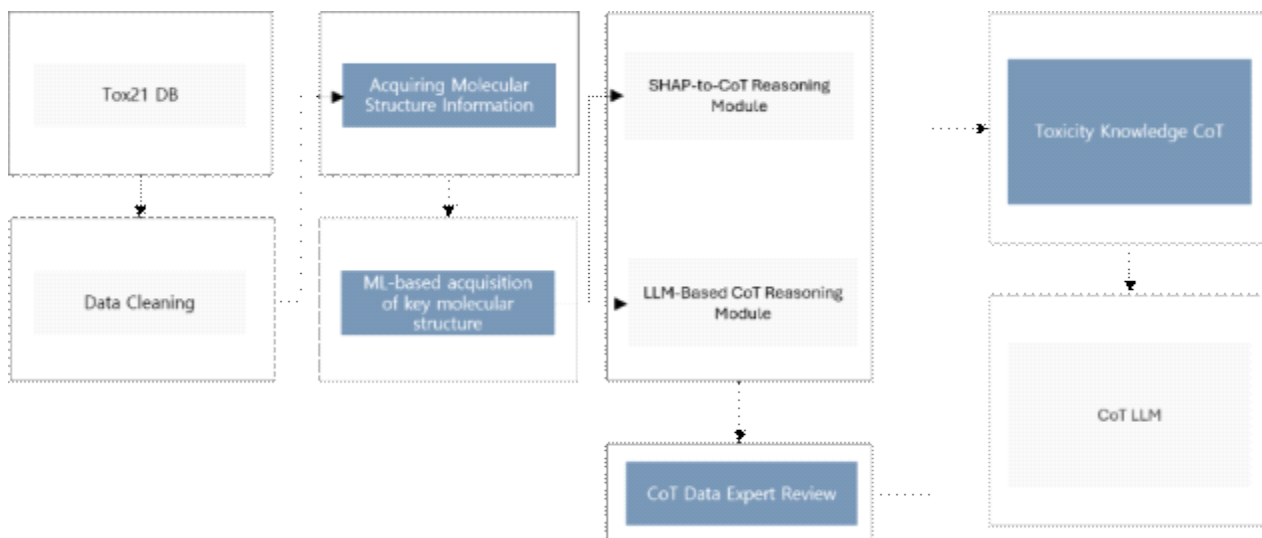
- 기존 CoT 기반 독성 예측 모델은 LLM이 직접 Chain-of-Thought를 생성함으로써 분자 독성에 대한 추론을 가능하게 했지만, 이 reasoning이 구조적 근거(feature evidence)에 기반하지 못한다는 한계를 지님. 즉, 독성 유발 기전과 연결된 명확한 증거 없이 언어 모델이 서술을 생성하기 때문에 기전적 타당성과 신뢰성이 부족함. 또한 이를 검증하는 전문가 검토 과정도 포함되지 않음.
- SHAP 기반의 설명성을 도입한 연구들도 있으나, 설명이 수치적 feature importance 수준에 머물러 있음. 독성의 원인·구조적 기전 등을 자연어 reasoning으로 해석하는 기능은 제공하지 않음. 또한 descriptor 기반 예측 모델이므로 구조적 다양성과 기전적 설명에 한계를 가짐.

○ 연구 목표

- 이와 같은 기존 연구의 제한점을 극복하기 위해, 본 연구에서는 다음의 프레임워크를 구축하였음.
(1) SHAP 기반의 구조적 근거(evidence)를 먼저 확보, (2) 이를 rule-based reasoning으로 변환 (3) LLM이 이를 확장해 자연어 CoT 생성 구조를 설계. (4) 전문가 검수(expert-in-the-loop) 시행. 이러한 프레임워크를 통해 생성된 reasoning의 기전적 정확성과 생물학적 타당성을 보정함으로써 기존 모델이 제공하지 못한 근거 기반·고신뢰 독성 추론 체계를 완성하고자 하였음.

□ 주요 설계

○ Architecture 구성도



○ 수도 코드 (Pseudo code)

Input : Tox21 toxicology dataset

Output : Toxicity Knowledge CoT dataset (expert-validated reasoning data)

#1. Acquire Molecular Structure Information

```
mol_structures ← get_molecular_structure(clean_data)
```

#2. ML-Based Extraction of Key Molecular Features

```
ml_model ← train_baseline_model(clean_data, mol_structures)
```

```
key_features ← extract_key_features(ml_model, mol_structures)
```

#3. SHAP-to-CoT Reasoning Module

```
shap_values ← compute_shap(ml_model, mol_structures)
```

```
cot_shap ← generate_cot_from_shap(shap_values)
```

#4. LLM-Based CoT Reasoning Module

```
cot_llm ← generate_cot_using_llm(key_features, mol_structures)
```

#5. Expert Review Module

```
cot_validated ← expert_review(cot_shap, cot_llm)
```

#6. Build Final Toxicity Knowledge CoT Dataset

```
toxicity_cot ← assemble_dataset(clean_data,  
                                mol_structures,  
                                cot_validated)  
  
return toxicity_cot
```

□ 기술 개발 결과

○ 주요 성능 지표(Performance Metrics)

	MMLU Accuracy	
Qwen	baseline	Toxicity Knowledge CoT
	89.6%	90.9%

□ 소스코드 공개

- Git-Hub 공개: <https://github.com/KwangSun-Ryu/ADMET-AGI-Knowledge-COT-and-Organoid-Source->

main

ADMET-AGI-Knowledge-COT-and-Organoid-Source- / Toxicity Knowledge CoT

Go to file

OREJH [강원대학교] 폴더 구조 변경

ab52be8 · 1

Name	Last commit message
..	
__pycache__	[강원대학교] 폴더 구조 변경
README.md	[강원대학교] 폴더 구조 변경
Toxicity Knowledge CoT 구현 결과서.pdf	[강원대학교] 폴더 구조 변경
evaluate.py	[강원대학교] 폴더 구조 변경
main.py	[강원대학교] 폴더 구조 변경
pipeline.sh	[강원대학교] 폴더 구조 변경
scoring.py	[강원대학교] 폴더 구조 변경
utils.py	[강원대학교] 폴더 구조 변경

README.md

Toxicity Knowledge CoT Framework

This repository provides a batch inference and evaluation pipeline designed to support toxicity knowledge Chain-of-Thought (CoT) evaluation.

The framework sends JSON-formatted samples to the server, collects model responses containing structured reasoning, and evaluates binary toxicity classification accuracy.

☐ 참고문헌

1.