Agentic AI for Materials Design

박찬영

Associate Professor, KAIST

Industrial and Systems Engineering
Graduate School of Data Science
Graduate School of AI

cy.park@kaist.ac.kr

Brief Bio

Chanyoung Park
Ph.D. in CSE POSTECH
Associate Professor
ISE KAIST
GSDS/GSAI KAIST

Contact Information
- cy.park@kaist.ac.kr
- http://dsail.kaist.ac.kr/

- 연구실: Data Science and Artificial Intelligence Lab
  - 총원 24명 (박사과정 17명, 석사과정 7명)

DSAIL Data Science & Artificial Intelligence

- Research Interest
  - AI for Science
    - Materials & Chemistry (Spectroscopy, Property Prediction, Retrosynthesis, Drug Discovery)
    - Biology (Single-cell Foundation Models, Virtual Cell, Protein, Perturbation)
    - Engineering (EDA, Circuit Analysis)
  - Agentic AI
    - AI agents for scientific reasoning and discovery, Tool-augmented LLM agents
  - Foundation AI Methods
    - Large Language Models & Multimodal Learning, Graph Neural Networks & Data Mining
  - Data-Centric AI
    - Robustness & Adversarial Learning, Imbalanced / Few-shot Learning, Explainable AI
- Professional Experience
  - Assistant/Associate Professor, **KAIST** (2020.11 – Present)
  - Visiting Scholar, **University of California San Diego**, Computer Science (2024.1 – 2025. 1)
  - Postdoc, **University of Illinois at Urbana-Champaign**, Computer Science (2019. 1 – 2020. 10)

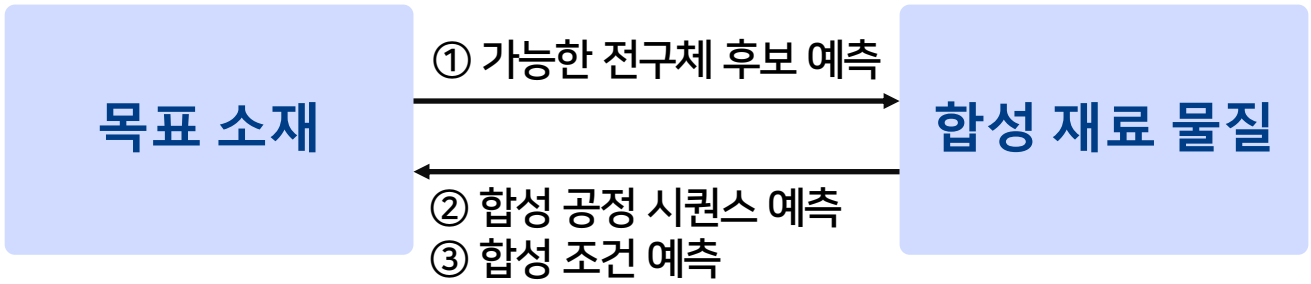
Outline

- 소재 합성 모델링
- 분광 데이터 분석
- Agentic AI for Science
- Fragmentation 기반 Molecule Representation Learning

Outline

- 소재 합성 모델링
- 분광 데이터 분석
- Agentic AI for Science
- Fragmentation 기반 Molecule Representation Learning

소재 합성 파이프라인



① 가능한 전구체 후보 예측
→ Retrieval-retro (NeurIPS 2024)

① 가능한 전구체 후보 예측
② 합성 공정 시퀀스 예측
→ MSP-LLM (Under Review)

예시)

- 목표소재: $Mg_{0.5}Zn_{0.5}Fe_2O_4$ (마그네슘-아연 페라이트(Mg-Zn ferrite))
- ① 합성재료물질 (전구체): Fe_2O_3, $MgCO_3$, ZnO
- ② 합성 공정 시퀀스: 건조→혼합→가열→혼합→성형→가열→어닐링→급랭
- ③ 합성조건: 각 공정 단계마다 온도/시간/압력 등

단계	온도	시간	핵심 포인트
건조	100–120°C	1–2 h	수분 제거
1차 가열	750–850°C	2 h	$MgCO_3$ 분해
2차 가열	950–1050°C	3–5 h	스피넬 형성
어닐링	800–900°C	1–3 h	결정 안정화
압력	100–300 MPa	—	치밀도 확보

Retrieval-Retro: Retrieval-based Inorganic Retrosynthesis with Expert Knowledge

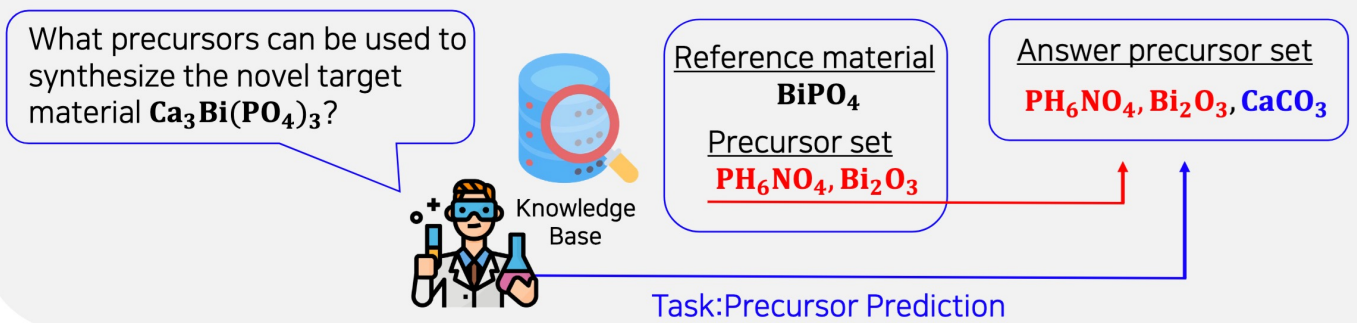
NeurIPS 2024

Heewoong Noh, Namkyeong Lee, Gyung S. Na, Chanyoung Park

Introduction

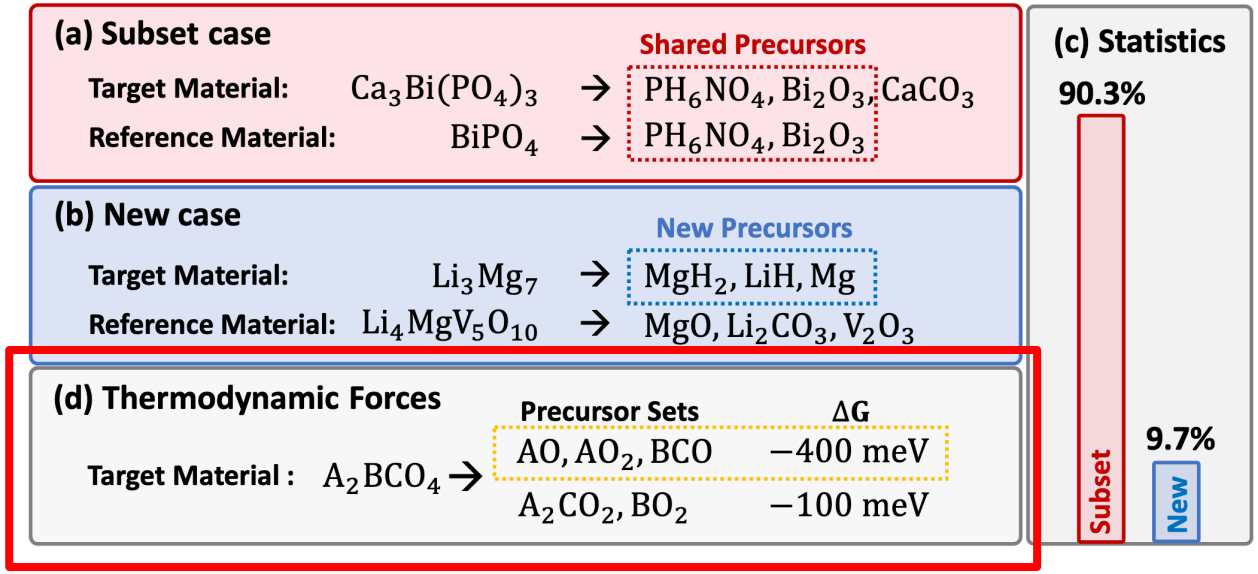
- 기존 소재 합성 접근 방식
  - Step 1. **신규 타겟 소재 정의** (새롭게 발견되거나 설계된 소재)
  - Step 2. **유사 소재 탐색:** DB에서 타겟과 유사한 기존 소재 검색
  - Step 3. **전구체 추론:** 참고(reference) 소재의 전구체 정보를 기반으로 타겟 소재의 전구체 예측
  - Step 4. **합성 레시피 최적화 및 실험 수행:** 실험 반복을 통한 공정 조정 및 성능 개선
- 기존 방식의 한계
  - DB내 **유사 소재 탐색에 의존**
  - **연구자의 경험과 직관에 크게 의존**
  - 반복 실험 중심의 시간·비용 소모적 과정
- 무기물은
  - 원자 수가 많고, 금속/다양한 원소 포함
  - DFT 계산도 불안정
  - 구조를 사전에 알기 어렵고 crystal structure prediction 비용이 매우 높음

→ 구조 대신 조성(composition) 기반 접근이 필요



Organic (유기물)	Inorganic (무기물)
분자 구조(SMILES, graph) 기반	구조 계산이 매우 비쌈
반응 메커니즘 이론이 비교적 정립	일반화된 합성 이론 부족
ML 연구 활발	ML 연구 상대적으로 적음

Motivation



도메인 지식 반영

- (a) Target이 Reference의 precursor 일부를 공유하는 경우
- (b) Target이 Reference와 precursor를 전혀 공유 X

→ 기존 레시피에 의존하면 절대 찾을 수 없음

- (c) 10%가 실제 혁신적인 합성 경로일 가능성이 높음
- (d) 열역학적 구동력 ΔG: 값이 클수록 합성가능성 높음

→ 기존연구는 precursor 유사성만 보고 ΔG 고려 안 함

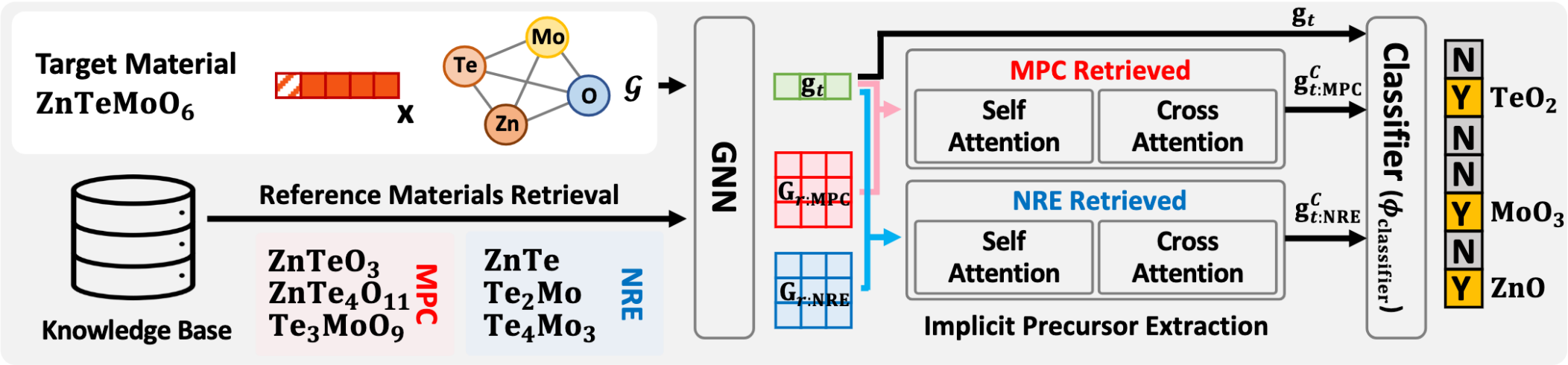
*깁스 자유에너지 $\Delta G \approx \Delta H = H_{Target} - H_{Precursor\ set}$

- $H_{target} < H_{precursor}$ (ΔG < 0) → 반응이 자발적으로 진행 가능
- ΔG가 더 음수 → 합성 가능성 더 높음

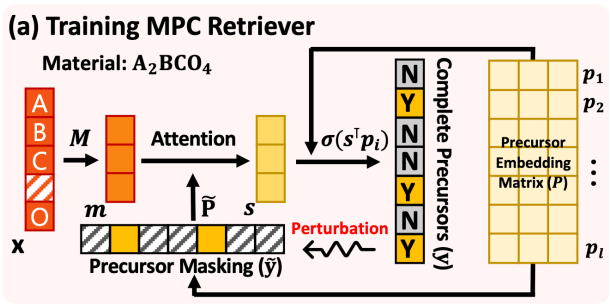
Retrieval-retro
- 기존 precursor를 단순 복사하지 않으면서
- domain knowledge (thermodynamics)를 반영하고
- 새로운 합성 레시피를 찾을 수 있는 방법

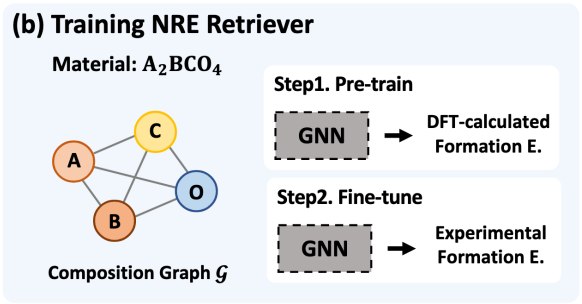
Method

MPC: Masked Precursor Completion
NRE: Neural Reaction Energy



- 1. 어떤 reference material을 가져올 것인가?





"비슷한 precursor 조합을 쓸 가능성이 높은" 물질을 찾음

물리적 반응 가능성 (ΔG) 기준 retrieval

- 2. precursor 정보를 attention을 통해 "암묵적으로" 추출

$$\mathbf{G}'^0_r = [\mathbf{g}'^1_r, \ldots, \mathbf{g}'^K_r] \qquad \mathbf{g}'^k_r = \phi_1(\mathbf{g}^k_r || \mathbf{g}_t)$$

$$\mathbf{G}'^s_r = \text{Self-Attention}(\mathbf{Q}_{\mathbf{G}'^{s-1}_r}, \mathbf{K}_{\mathbf{G}'^{s-1}_r}, \mathbf{V}_{\mathbf{G}'^{s-1}_r}) \in \mathbb{R}^{K \times D}$$

$$\mathbf{g}^c_t = \text{Cross-Attention}(\mathbf{Q}_{\mathbf{g}^{c-1}_t}, \mathbf{K}_{\mathbf{G}'^S_r}, \mathbf{V}_{\mathbf{G}'^S_r}) \in \mathbb{R}^D$$

- precursor 자체를 직접 쓰지 않고 representation 공간에서 정보 추출
- 모델이 기존 precursor에 갇히지 않음

→ latent space에서 새로운 조합 가능

Experiments: Dataset & Evaluation Protocol

- **Data Source (대규모 실제 문헌 기반 synthesis 데이터 (Kononova et al., 2019))**
  - 33,343 inorganic synthesis recipes
  - 24,304 materials science papers에서 텍스트 마이닝으로 추출
  - 기존 연구에서 구축된 대규모 문헌 기반 합성 데이터셋 / 무기 precursor 예측을 위한 대표 벤치마크
- **Knowledge Base 구성**
  - Training set을 knowledge base로 활용
  - Retrieval 기반 방법에서 → 학습 데이터 내 reference materials 검색
- **Evaluation Splits**
  - Random Split (80/10/10)
  - Year Split
    - Train: ~2014
    - Valid: 2015–2016
    - Test: 2017–2020

      → 과거 데이터로 학습하여 미래에 등장한 물질의 합성 경로 예측

      → 더 현실적이고 어려운 평가 환경

Table 6: The distribution of data based on the number of ground truth precursor sets.

# Precursor sets	1	2	3	4	5	6	7	8	9	10	11	12	13	14	17	23	28	29	34	Total
# Data	26,944	1,168	257	108	42	32	13	8	7	5	4	1	2	2	1	1	1	1	1	28,598

<material 당 가능한 precursor set 개수 분포>

(Kononova et al., 2019) Text-mined dataset of inorganic materials synthesis recipes (Nature Scientific Data 2019)

Result

Model	Interaction	Retrieval	(a) Year Split						(b) Random Split					
			Top-K Accuracy				Recall		Top-K Accuracy				Recall	
			Top-1	Top-3	Top-5	Top-10	Macro	Micro	Top-1	Top-3	Top-5	Top-10	Macro	Micro
Composition MLP	✗	✗	31.60 (1.70)	34.37 (1.58)	35.22 (1.43)	36.56 (0.160)	31.42 (0.030)	31.44 (0.060)	58.56 (0.47)	62.20 (0.36)	62.95 (0.029)	64.32 (0.42)	54.56 (0.44)	55.35 (0.57)
He et al. [12]	✗	✓	45.03 (1.85)	48.02 (1.86)	49.11 (1.77)	51.09 (1.93)	44.72 (1.83)	44.75 (1.86)	61.94 (1.5)	66.44 (1.48)	67.46 (1.55)	68.84 (1.65)	58.55 (1.45)	59.35 (1.34)
ElemwiseRetro	✓	✗	53.45 (0.58)	57.07 (0.52)	58.19 (0.72)	60.84 (0.78)	53.12 (0.60)	53.19 (0.60)	77.23 (0.70)	80.93 (0.54)	81.57 (0.67)	82.78 (0.64)	72.33 (1.14)	73.26 (0.99)
Roost	✓	✗	54.38 (0.75)	57.82 (0.81)	58.82 (1.00)	60.71 (1.15)	54.01 (0.75)	54.04 (0.74)	78.42 (0.91)	82.32 (0.91)	83.07 (0.83)	84.10 (0.66)	73.38 (1.41)	74.46 (1.22)
CrabNet	✓	✗	57.15 (0.77)	61.60 (0.85)	62.44 (0.82)	64.14 (0.86)	56.79 (0.77)	56.82 (0.77)	78.69 (0.78)	81.62 (0.74)	82.27 (0.67)	83.35 (0.56)	73.27 (1.21)	74.28 (0.99)
Graph Network	✓	✗	58.95 (0.41)	63.10 (0.63)	64.07 (0.68)	66.30 (0.62)	58.54 (0.42)	58.61 (0.41)	77.91 (1.31)	81.55 (0.98)	82.37 (0.92)	83.50 (0.90)	72.96 (1.53)	73.88 (1.29)
Graph Network + MPC	✓	✓	60.01 (1.10)	64.15 (1.10)	65.15 (1.17)	67.19 (0.83)	59.61 (1.10)	59.66 (1.10)	79.09 (1.25)	82.95 (1.13)	83.82 (1.19)	84.97 (0.94)	73.86 (1.34)	74.81 (1.23)
Retrieval-Retro	✓	✓	**61.16** (0.38)	**65.92** (0.71)	**67.18** (0.76)	**69.45** (1.03)	**60.97** (0.62)	**61.06** (0.62)	**79.81** (0.68)	**83.62** (0.77)	**84.46** (0.78)	**85.70** (0.88)	**74.61** (0.98)	**75.49** (0.89)

- **Year split는 더 어렵지만 현실적인 세팅:** 전체적으로 random split 대비 성능이 낮으며, 실제 "미래 소재 예측" 상황을 더 잘 반영.
- **Retrieval 기반 모델이 Year split에서 특히 중요:** reference material을 활용하는 모델들이 non-retrieval 모델보다 일관되게 높은 성능을 보임.
- **Retrieval-Retro가 Year split에서 가장 큰 개선:** Top-1/3/5/10 정확도와 Macro/Micro Recall 모두 최고 성능을 기록하며, 특히 Graph Network + MPC 대비 명확한 추가 향상을 달성.

Result

GNN(target composition + retrieved precursor)

Model	Refer.	Retriever		Subset Case				New Case			
		MPC	NRE	Top-1	Top-3	Top-5	Top-10	Top-1	Top-3	Top-5	Top-10
Graph Network	Explicit	✗	✗	63.98 (0.34)	67.95 (0.53)	68.83 (0.64)	70.83 (0.81)	16.37 (1.91)	22.00 (3.47)	23.78 (3.48)	27.93 (1.66)
		✓	✗	65.01 (1.10)	69.06 (1.17)	69.98 (1.22)	72.03 (0.92)	17.63 (1.66)	22.45 (2.63)	24.22 (2.65)	26.22 (2.74)
Retrieval-Retro	Implicit	✓	✗	65.07 (0.80)	69.44 (1.27)	70.41 (1.24)	72.47 (1.46)	19.70 (1.08)	24.52 (1.42)	26.30 (1.28)	30.15 (2.05)
		✓	✓	**66.00** (0.32)	**70.51** (0.61)	**71.76** (0.61)	**73.92** (0.90)	**20.15** (1.29)	**27.04** (1.93)	**28.37** (2.05)	**31.56** (3.44)

- Explicit 방식(Graph Network + MPC)은 subset case에서는 성능이 향상되지만, **new case에서는 성능 개선이 제한적이거나 감소**하여 새로운 합성 경로 발견에 취약함.
- Implicit 방식(Retrieval-Retro)은 subset과 new case 모두에서 일관된 성능 향상, 특히 **new case에서 큰 개선**을 보임.
- **MPC + NRE를 함께 사용할 때 최고 성능**, 두 retriever가 상호보완적으로 작용함.

Result

Table 4: Qualitative Analysis (Target Material: $Pb_9[Li_2(P_2O_7)_2(P_4O_{13})_2]$).

Model	Retriever	Retrieved Material	Corresponding Precursor Sets	Predicted Precursor Set	Answer
Only MPC	MPC	$LiNaPbPO$ $Li_{0.5}Na_{0.5}PO_3$ $Li_3V_{1.92}Al_{0.08}(PO_4)_3$	$\{Li_2CO_3, H_3PO_4, Na_2CO_3, Pb_3O_4\}$ $\{Li_2CO_3, NH_4H_2PO_4, NaPO_3\}$ $\{Al, V_2O_5, LiH_2PO_4\}$	$\{Li_2CO_3, NH_4H_2PO_4\}$	✗
MPC + NRE (Ours)	MPC	$LiNaPbPO$ $Li_{0.5}Na_{0.5}PO_3$ $Li_3V_{1.92}Al_{0.08}(PO_4)_3$	$\{Li_2CO_3, H_3PO_4, Na_2CO_3, Pb_3O_4\}$ $\{Li_2CO_3, NH_4H_2PO_4, NaPO_3\}$ $\{Al, V_2O_5, LiH_2PO_4\}$	$\{Li_2CO_3, NH_4H_2PO_4, PbO\}$	✓
	NRE	$Pb_3(PO_4)_2$ Li_3P PbP_7	$\{PbO, NH_4H_2PO_4\}$ $\{P, Li\}$ $\{P, Pb\}$		

- MPC만 쓰면 precursor를 일부만 맞추지만,
- MPC + NRE를 같이 쓰면 전체 precursor set을 정확히 예측

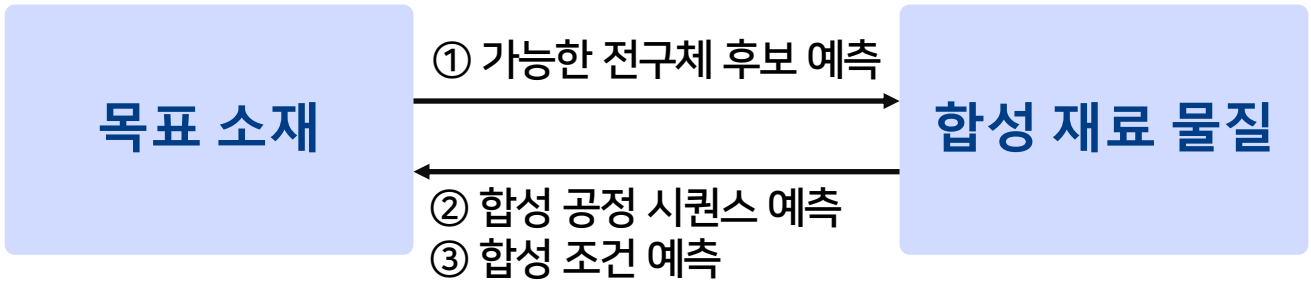
MSP-LLM: A Unified Large Language Model Framework for Complete Material Synthesis Planning

Under Review

Published at “AI4Mat: AI for Accelerated Materials Design (ICLR 2026 Workshop)”

Heewoong Noh, Gyoung S. Na, Namkyeong Lee, Chanyoung Park

소재 합성 파이프라인



① 가능한 전구체 후보 예측
→ Retrieval-retro (NeurIPS 2024)

① 가능한 전구체 후보 예측
② 합성 공정 시퀀스 예측
→ MSP-LLM (Under Review)

예시)

- 목표소재: $Mg_{0.5}Zn_{0.5}Fe_2O_4$ (마그네슘-아연 페라이트(Mg-Zn ferrite))
- ① 합성재료물질 (전구체): Fe_2O_3, $MgCO_3$, ZnO
- ② 합성 공정 시퀀스: 건조→혼합→가열→혼합→성형→가열→어닐링→급랭
- ③ 합성조건: 각 공정 단계마다 온도/시간/압력 등

단계	온도	시간	핵심 포인트
건조	100–120°C	1–2 h	수분 제거
1차 가열	750–850°C	2 h	$MgCO_3$ 분해
2차 가열	950–1050°C	3–5 h	스피넬 형성
어닐링	800–900°C	1–3 h	결정 안정화
압력	100–300 MPa	—	치밀도 확보

Overview

- Material Synthesis Planning (MSP)

$$f : \mathcal{M} \rightarrow 2^{\mathcal{M}} \times \mathcal{O}^*$$

target material → (precursor set, synthesis operation sequence)

$La_{1.2}Ca_{1.8}Mn_2O_7$ → ({$CaCO_3$,La_2O_3,MnO_2}, (mixing → heating → mixing → sintering))

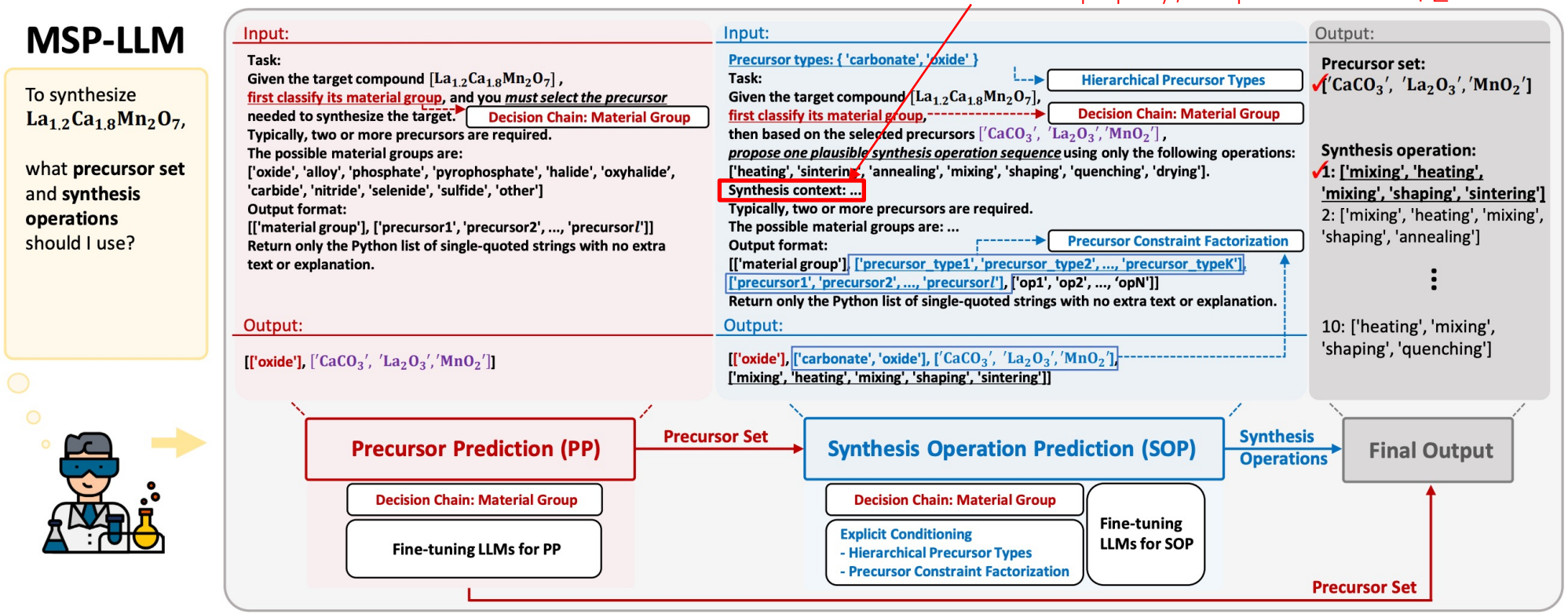
- 기존 연구: Precursor Prediction (PP) or Synthesis Operation Prediction (SOP) 중 한가지 Task만 집중
- 본 연구: LLM 기반으로 전체 MSP를 end-to-end로 푸는 통합 프레임워크를 제안

- 왜 LLM인가?
  - 모든 입력은 text (chemical formula) / LLM은 대규모 물리/화학 지식 학습 / reasoning capability 보유
  - 하지만, 아직 MSP에 특화된 fine-tuning 전략이 없다

Method

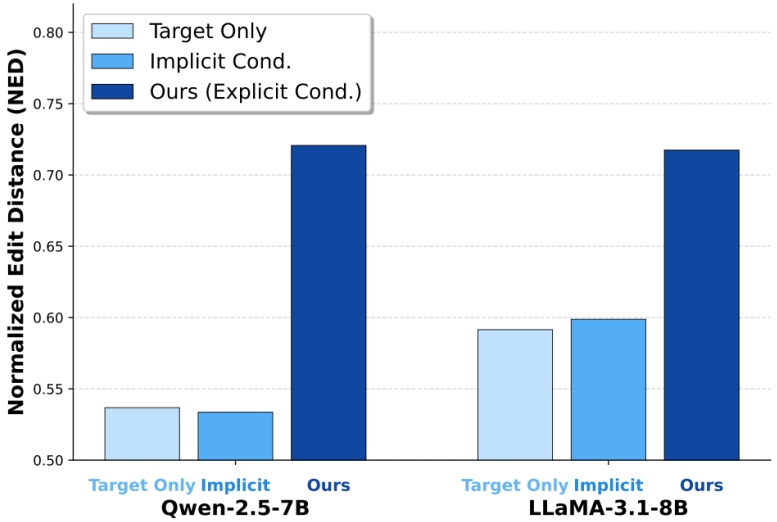
- MSP를 2단계로 분해: 1) Precursor Prediction (PP), 2) Synthesis Operation Prediction (SOP)
- Material Group을 넣어서 Decision chain으로 모델링 (실제로 Material group별로 합성 전략이 다름)
- LLM이 바로 생성하지 않고, 1) 먼저 material group을 예측 → 2) 그 group에 조건화해서 생성
  → search space를 줄이며 chemically plausible subspace로 제한

논문 제목으로부터 'host material', 'dopant or substitution', 'material class', 'functional property', 'composition control' 추출



Explicit conditioning

- 관측: Precursor를 input에 넣는다고 LLM이 잘 활용하지 않는다.
  - Target-only: Task description + Target material
  - Implicit conditioning: Task description + Target material + Precursor set



- Precursor set을 input prompt에 추가한다고 해서 성능 향상 X
- Why?
  - precursor formula들이 길고 복잡함 / 여러 chemical formula가 연속으로 나열됨
  - LLM의 decoder state가 이를 안정적으로 유지하지 못함

  → LLM은 다음 토큰 예측만 잘하면 되기 때문에, precursor 정보를 안 써도 loss를 줄일 수 있으면 그냥 무시

- **Ours (Explicit conditioning)**: Task description + Target material + Precursor set + Precursor Types
  - Synthesis-relevant behavior를 직접 encoding (carbonate → heating 필요 등)
  - Type 정보 없이는 개별 물질에 대한 operation을 배워야 하지만, Type 정보가 있으면 한번만 배우면 됨
  - Inductive bias (예.$CaCO_3$가 carbonate라는 것만 알면 → 새로운 carbonate에도 일반화 가능)
  - Decoder state 안정성: 긴 chemical string보다 유지 쉬움

```
P = {CaCO3, La2O3, MnO2}
Types = {carbonate, oxide}
```

Precursor Conditioned Factorization (PCF)

- Precursor 정보를 단순 input context로 넣는 것이 아니라, 모델이 먼저 precursor 정보를 예측하도록 만들어 operation 생성이 precursor에 의존하도록 강제하는 방법
- SOP task를 다음과 같이 정의

Synthesis operation prediction (SOP)

$$P(G, Z, O \mid X) = P(G \mid X)P(Z \mid X, G)P(O \mid X, G, Z)$$

Material group 예측

Precursor 정보를 사용하도록 강제

Z: (Precursor type, Precursor set)
G: Material Group

  - 반면, Implicit conditioning: $P(O|X, Z), Z =$ Precursor set
    - 이 경우 모델이 Precursor Z를 보지 않고 target material만 보고 operation을 예측을 해도 Loss가 줄어듦
  - PCF의 경우 1) precursor Z를 예측해야 하고, 2) 그걸 기반으로 operation을 생성 해야해서 Z를 무시할 수 없음

Experiments: Setup

- Dataset
  - Inorganic material synthesis dataset (합성 레시피를 text mining으로 추출)
  - 10,851 synthesis records 사용
  - 각 샘플: (target material, precursor set, synthesis operations)
  - Precursor: 빈도 ≥ 5 인 precursor만 사용 → 288 unique precursors
  - Synthesis operations: 총 7 types {mixing, drying, heating, sintering, annealing, quenching, shaping}
- Model Backbones: Fine-tuned open-source LLMs: LLaMA-3.1-8B / Qwen-2.5-7B
- Training setup
  - QLoRA parameter-efficient fine-tuning / PP와 SOP 각각 독립적으로 fine-tuning
  - 최종 MSP는 PP → SOP 두 단계 파이프라인으로
- Evaluation Protocol - Dataset Split: 두 가지 split 사용 (Train/val/test = 8:1:1)
  - (1) Random split: 동일 target material이 train/test에 모두 등장 가능. 단, precursor / operation sequence는 다름
  - (2) Target-disjoint split: target material을 완전히 분리 → unseen material generalization 평가

Result: Precursor Prediction (PP) & Sequence Operation Prediction (SOP)

- NED (Normalized Edit Distance): 예측된 operation sequence를 GT sequence로 바꾸는 데 필요한 edit operation 수를 길이로 정규화한 거리 기반 유사도
- LCS (Longest Common Subsequence): 예측과 GT 사이에서 순서를 유지한 채 공통으로 등장하는 operation subsequence의 최대 길이
- F1 (operation multiset overlap): operation 종류와 개수의 겹침 정도를 precision-recall 기반 F1 score로 측정 (순서는 무시)

Table 1. PP performance on two dataset splits.

Method	Split 1				Split 2			
	Top-1	Top-3	Top-5	Top-10	Top-1	Top-3	Top-5	Top-10
Specialized Models								
He et al. (2023)	53.11	60.05	61.06	63.07	54.01	61.10	62.69	65.02
Retrieval-Retro (Noh et al., 2024)	60.88	69.38	70.93	74.13	66.32	74.16	75.84	78.54
Large Language Models (API)								
GPT-4o-mini (20-shot)	56.58	65.36	67.82	70.66	54.85	63.40	67.91	70.34
GPT-4o-mini (40-shot)	55.21	63.44	66.27	69.10	55.97	64.27	67.82	71.36
GPT-4o (20-shot)	60.05	68.56	71.76	75.05	60.17	69.31	73.04	76.31
GPT-4o (40-shot)	61.79	70.48	73.40	76.32	61.57	71.27	73.97	77.05
GPT-5.1 (20-shot)	64.72	77.42	80.53	83.27	67.07	77.61	82.37	85.17
GPT-5.1 (40-shot)	64.08	76.33	79.62	83.46	66.60	77.89	82.74	86.01
Claude Haiku 4.5 (20-shot)	55.03	68.37	74.50	78.88	56.44	68.84	73.13	78.92
Claude Haiku 4.5 (40-shot)	57.31	68.74	74.31	78.43	56.53	71.64	76.12	81.81
MSP-LLM (Qwen-2.5-7B)	**71.02**	85.37	90.49	94.06	71.92	88.62	92.72	95.53
MSP-LLM (LLaMA-3.1-8B)	70.75	**85.56**	**91.68**	**94.97**	**72.85**	**88.99**	**93.56**	**96.36**

Table 2. SOP performance on two dataset splits.

Method	Split 1				Split 2			
	Top-10	NED↑	LCS↑	F-1↑	Top-10	NED↑	LCS↑	F-1↑
Specialized Models								
Retrieval based	10.4	0.6072	0.7194	0.7596	11.19	0.6030	0.7198	0.7564
SPENDE (Na, 2023)	N/A	0.4508	0.5824	0.6013	N/A	0.4522	0.5821	0.6010
Large Language Models (API)								
GPT-4o-mini (20-shot)	5.85	0.5717	0.6992	0.7355	7.09	0.5847	0.7089	0.7407
GPT-4o-mini (40-shot)	5.94	0.5844	0.7092	0.7454	7.93	0.5934	0.7141	0.7442
GPT-4o (20-shot)	5.03	0.5654	0.6925	0.7219	3.82	0.5670	0.6865	0.7166
GPT-4o (40-shot)	4.34	0.5693	0.6890	0.7195	4.94	0.5749	0.6924	0.7245
GPT-5.1 (20-shot)	2.93	0.5334	0.6583	0.7101	2.33	0.5301	0.6547	0.6968
GPT-5.1 (40-shot)	2.74	0.5434	0.6660	0.7146	2.33	0.5261	0.6498	0.6925
Claude Haiku 4.5 (20-shot)	2.56	0.5534	0.6727	0.7174	1.68	0.5439	0.6656	0.7017
Claude Haiku 4.5 (40-shot)	4.11	0.5669	0.6801	0.7172	3.26	0.5569	0.6730	0.7037
MSP-LLM (Qwen-2.5-7B)	**26.51**	**0.7207**	**0.8100**	**0.8359**	**32.18**	0.7420	**0.8240**	0.8493
MSP-LLM (LlaMA-3.1-8B)	**26.51**	0.7174	0.8075	0.8353	31.44	**0.7424**	0.8235	**0.8498**

MSP-LLM이 모든 baseline 대비 크게 향상된 성능
→ 단순 prompting 기반 LLM이나 retrieval 모델보다 task-specific fine-tuning + material group decision chain이 precursor 예측에 효과적

MSP-LLM이 sequence generation 성능에서 압도적 우위
→ Explicit conditioning + precursor constraint factorization이 precursor 정보를 효과적으로 활용하게 만들어 operation sequence 예측 성능을 크게 향상

Result: Material Synthesis Planning (MSP)

Method	Split 1				Split 2			
	Top-1	Top-3	Top-5	Top-10	Top-1	Top-3	Top-5	Top-10
Large Language Models (API)								
GPT-4o-mini (20-shot)	0.0	0.09	0.09	0.19	0.19	0.28	0.28	0.47
GPT-4o-mini (40-shot)	0.0	0.0	0.09	0.09	0.27	0.27	0.37	0.64
GPT-4o (20-shot)	0.27	0.55	0.64	0.73	0.47	0.47	0.47	0.56
GPT-4o (40-shot)	0.27	0.27	0.37	0.64	0.37	0.84	0.93	1.12
GPT-5.1 (20-shot)	0.46	0.64	0.73	1.01	0.65	0.75	0.75	0.75
GPT-5.1 (40-shot)	0.55	0.64	0.64	0.73	0.37	0.37	0.47	0.47
Claude Haiku 4.5 (20-shot)	0.56	0.65	0.65	0.74	0.56	0.65	0.65	0.75
Claude Haiku 4.5 (40-shot)	0.56	0.65	0.65	0.74	0.37	0.84	0.93	1.12
GPT-5.1(PP) + GPT-4o-mini (SOP)	0.91	1.65	2.01	3.75	1.4	2.71	3.64	5.41
MSP-LLM (Qwen-2.5-7B)	**5.76**	**10.79**	**13.62**	**18.56**	6.62	**13.34**	**16.79**	22.48
MSP-LLM (LLaMA-3.1-8B)	4.75	9.69	13.35	18.19	**6.72**	12.12	16.60	**23.23**

* precursor set + operation sequence 둘 다 맞아야 correct

- 단일 API LLM은 거의 실패: 대부분 Top-10 정확도가 2% 미만

→ precursor와 operation을 동시에 맞추는 complete MSP는 단일 모델로 매우 어려움.

- Hybrid pipeline도 제한적 개선: PP에 강한 모델 + SOP에 강한 모델을 결합해도 Top-10 ≈ 3~5% 수준

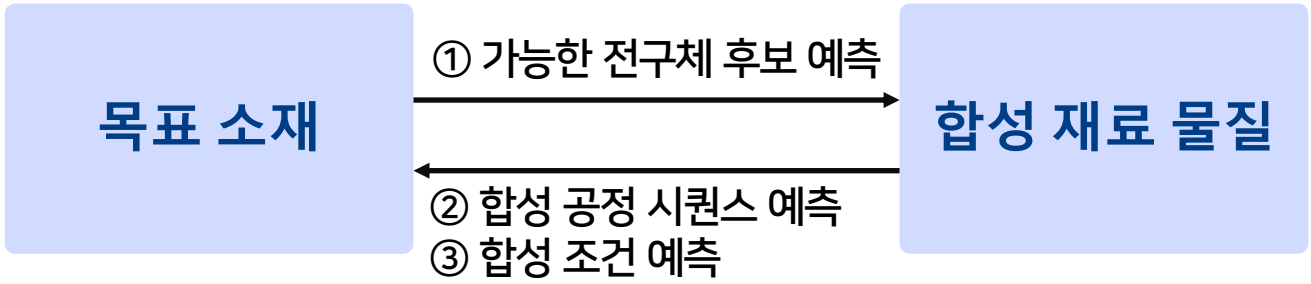
→ 두 단계 분리는 도움 되지만 여전히 낮은 성능.

- MSP-LLM이 압도적 우위: Top-10 18~23% 달성

→ 구조적 decision chain + explicit conditioning이 complete MSP에서 결정적 효과.

- 하지만, 전반적으로 낮은 성능

Future work



① 가능한 전구체 후보 예측
→ Retrieval-retro (NeurIPS 2024)

① 가능한 전구체 후보 예측
② 합성 공정 시퀀스 예측
→ MSP-LLM (Under Review)

① 가능한 전구체 후보 예측
② 합성 공정 시퀀스 예측
③ 합성 조건 예측
→ Ongoing work

예시)

- 목표소재: $Mg_{0.5}Zn_{0.5}Fe_2O_4$ (마그네슘-아연 페라이트(Mg-Zn ferrite))
- ① 합성재료물질 (전구체): Fe_2O_3, $MgCO_3$, ZnO
- ② 합성 공정 시퀀스: 건조→혼합→가열→혼합→성형→가열→어닐링→급랭
- ③ 합성조건: 각 공정 단계마다 온도/시간/압력 등

단계	온도	시간	핵심 포인트
건조	100–120°C	1–2 h	수분 제거
1차 가열	750–850°C	2 h	$MgCO_3$ 분해
2차 가열	950–1050°C	3–5 h	스피넬 형성
어닐링	800–900°C	1–3 h	결정 안정화
압력	100–300 MPa	—	치밀도 확보

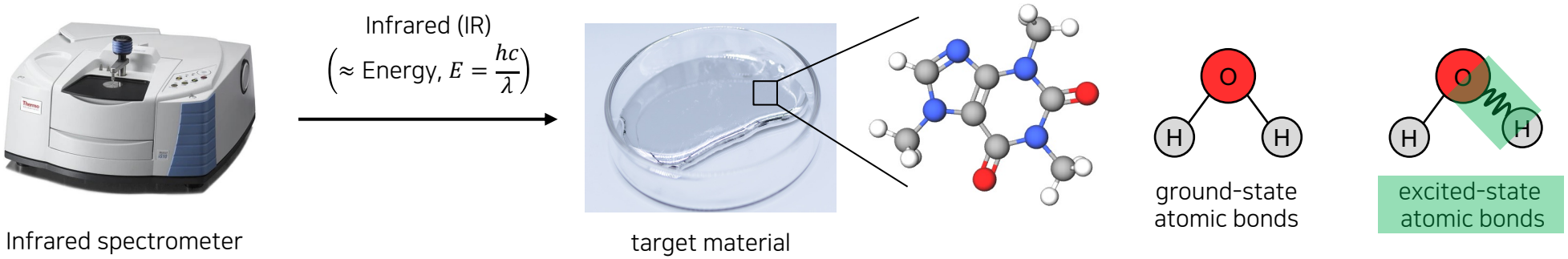
Outline

- 소재 합성 모델링
- 분광 데이터 분석
- Agentic AI for Science
- Fragmentation 기반 Molecule Representation Learning

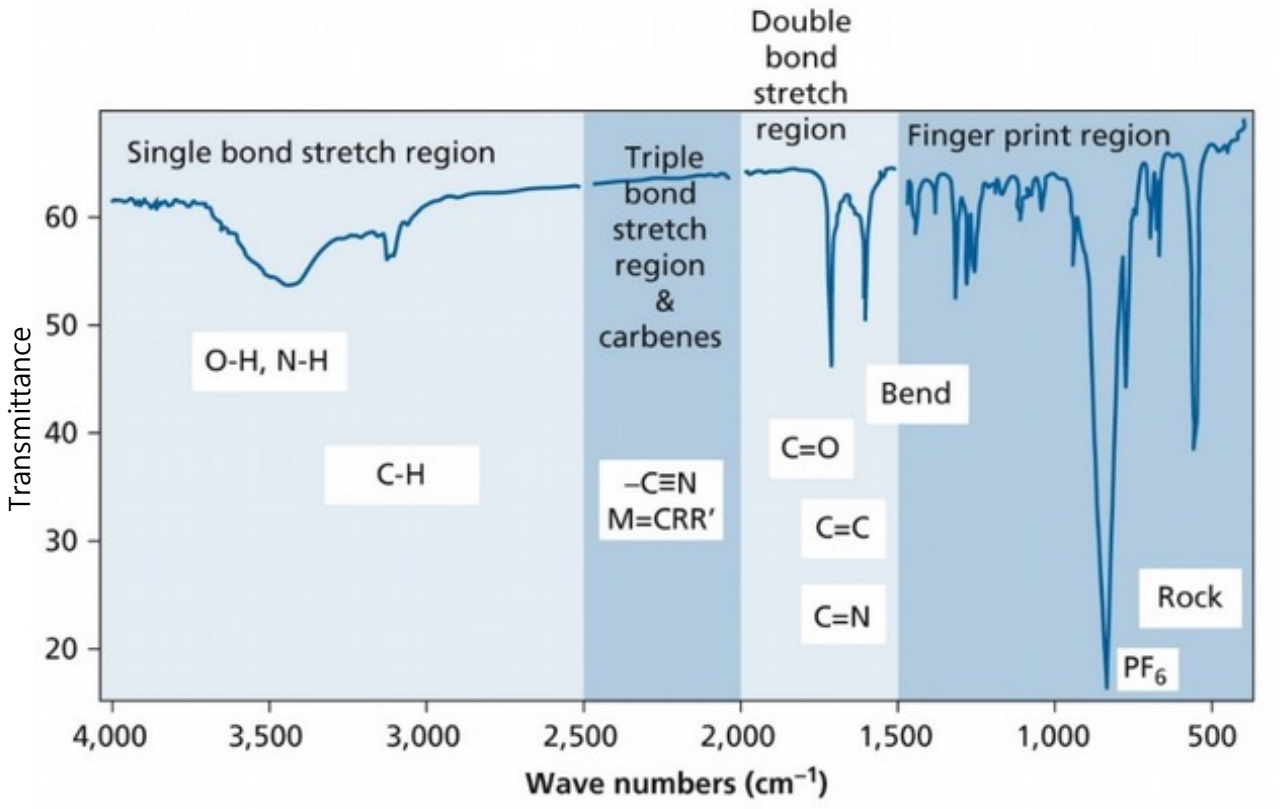
Spectroscopic data (분광 데이터)

- 분광 데이터는 물질과 에너지의 상호작용 결과로 얻어지는 신호
  - 빛(또는 다른 에너지)을 물질에 비췄을 때 물질이 어떻게 반응하는지 (어떤 파장에 반응하는지) 기록
  - 특정 파장에서의 스펙트럼 패턴에 원자·분자 구조 정보가 간접적으로 반영됨 (특정 원자 결합은 특정한 에너지 준위에서만 진동)
- 대표적 예
  - 적외선 분광(IR)
  - 라만 분광(Raman)
  - 핵자기공명(NMR-Nuclear magnetic resonance)
  - 질량분석(MS-Mass Spectrometry)

👉 보이지 않는 구조를 간접적으로 알아내기 위해 분광을 한다



IR Spectrum 분석 예시



BOND TYPE	$\bar{\nu}$	BOND TYPE	$\bar{\nu}$
RO-H (alcohol and phenols)	3,200–3,650	R-C(=O)-OR (esters; C=O)	1,735–1,750
R-C(=O)-O—H (carboxylic acids; O-H)	2,500–3,000	R-C(=O)-OH (carboxylic acids; C=O)	1,710–1,770
R_3C—H (alkane)	2,840–2,960	R-C(=O)-H; R-C(=O)-R (aldehyde and ketone C=O)	1,690–1,745
R(H)C=C(H)R (alkene; C-H)	3,050–3,100	C=C (alkenes)	1,620–1,680
RC≡C-H (alkyne)	3,280–3,310	RC≡CR (alkyne)	2,100–2,260
Ar–H (aromatic stretching) (bending)	~3,300 680–860	RC≡N	2,220–2,260

방법론별 특징 비교

구분	IR	Raman	NMR	MS
기본 원리	진동 흡수	진동 산란	핵 스핀 공명	이온 질량 측정
에너지 상호작용	적외선 흡수	레이저 산란	자기장 + RF	이온화 + 전기장
주는 정보	결합 종류	결합/대칭성	연결성, 환경	분자량, 조성
구조 해상도	작용기 수준	작용기 수준	원자 연결성	분자 조성
시료 상태	고체/액체/기체	대부분 가능	주로 액체	대부분 가능
파괴성	비파괴	비파괴	비파괴	보통 파괴적
장점	빠름, 저렴	물 영향 적음	구조 정보 풍부	정확한 분자량
한계	구조 완전 규명 어려움	형광 간섭	장비 고가	구조 정보 제한적

왜 분광 기반 구조 추론은 어려운가?

- **간접 관측 문제 (Inverse Problem)**
  - 구조를 직접 보는 것이 아님
  - 신호 → 구조로 역추론해야 함
- **구조적 모호성 (Ambiguity)**
  - 서로 다른 구조가 비슷한 스펙트럼을 만들 수 있음
  - 단일 peak는 구조를 결정하지 못함
  - 전체 패턴을 종합적으로 해석해야 함
- **신호의 집합성 (Aggregation)**
  - Spectrum은 여러 진동 모드 신호의 합
  - 환경 및 실험 조건에 따른 변화 포함
  - Noise 및 측정 오차 존재

그렇다면 우리는 분광 데이터를 AI로 어떻게 모델링해야 할까?

👉 **강한 Inductive bias가 필요**

1. 전문가의 해석 과정을 모델링 하자 → IR-Agent (ICLR 2026)
   - Spectrum은 전문가가 읽는 신호 (local + global pattern)
   - Reasoning 과정을 학습하자

2. Spectrum의 생성 메커니즘을 모델링 하자 → PCSP (Under Review)
   - 여러 진동 모드의 신호가 중첩된 결과
   - 스펙트럼의 생성 과정(generative process)을 모델링하자

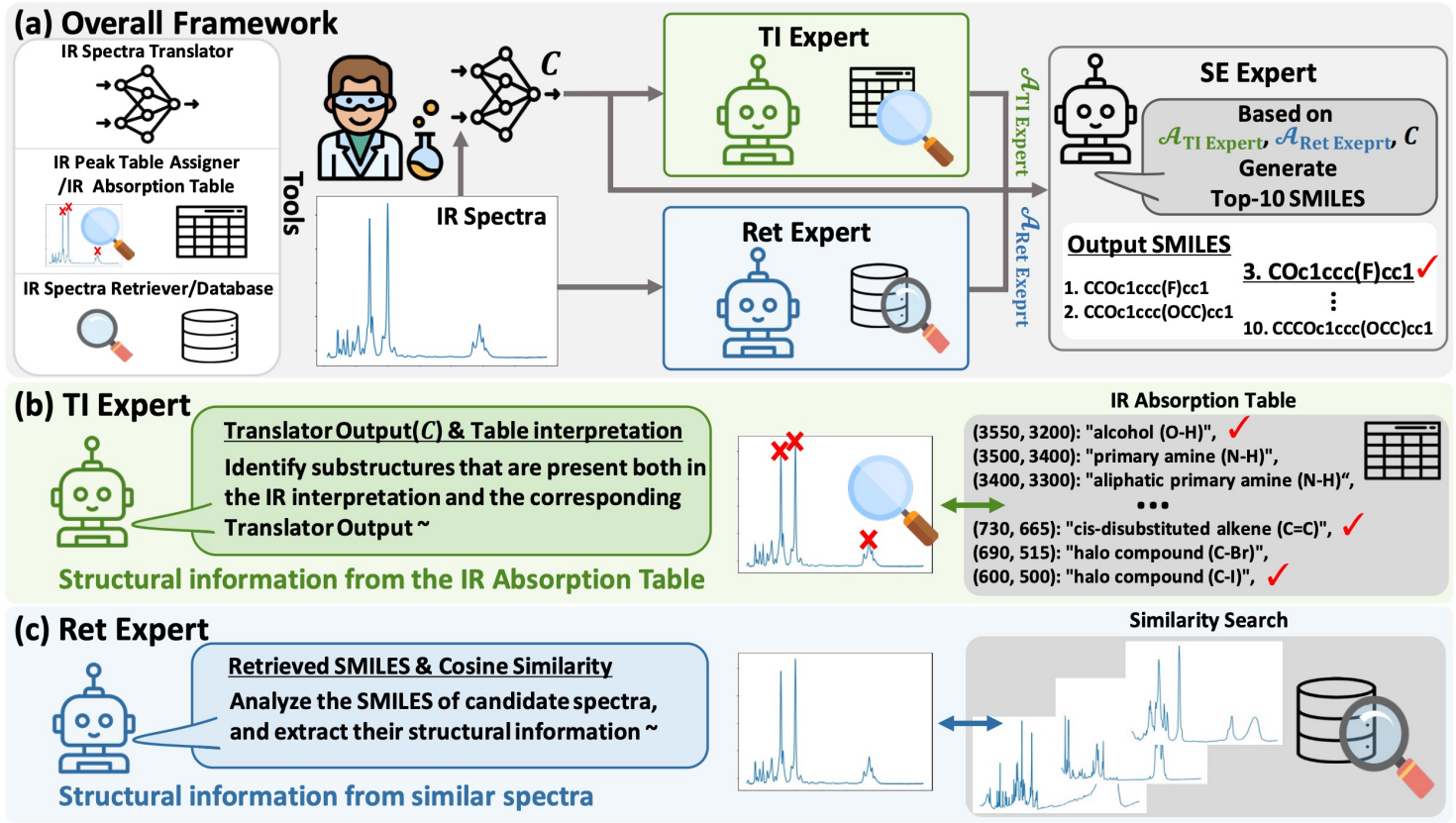
IR-Agent: Expert-Inspired LLM Agents for Structure Elucidation from Infrared Spectra

ICLR 2026

Heewoong Noh, Namkyeong Lee, Gyoung S. Na, Kibum Kim, Chanyoung Park

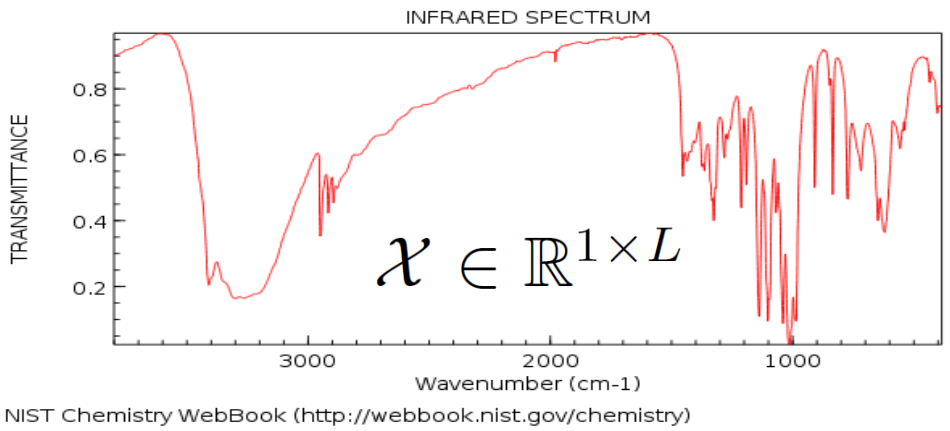
IR-Agent - Motivation

- Spectrum은 전문가가 해석하는 신호
- 단일 peak가 아니라 조합적 해석 필요
- **핵심 아이디어**
  - Substructure-level reasoning
  - Multi-agent 협업 추론



문제 정의

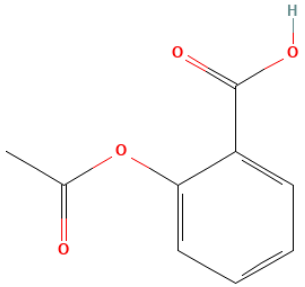
IR spectrum of a molecule



ML model

Output: SMILES (Full molecular structure)

CC(=O)OC1=CC=CC=C1C(=O)O



- **Task:** 미지 물질의 IR 스펙트럼으로 (Input: X) 부터 분자 구조 (Output: SMILES)를 예측하는 문제
- **기존 접근 방법**
  - 화학자들이 IR 스펙트럼의 피크를 수동으로 식별하고, IR absorption table을 기반으로 해석 (경험과 직관에 의존)
  - 대부분의 머신러닝 접근은 작용기 (functional group) 검출 수준에 한정됨
  - 많은 기존 방법은 Transformer 기반의 sequence-to-sequence 모델링에 의존
- **왜 어려운가?** 스펙트럼 특징이 복잡하고 서로 겹쳐 있어 해석이 어려움

IR Spectrum 분석을 위한 LLM 기반 Agent 프레임워크의 필요성

- IR 스펙트럼 분석은
  - 인간 전문가가 사용하는 분석 과정을 통합적으로 반영하는 것이 중요
  - 전문가들은 IR absorption table을 해석하여 피크 위치로부터 local substructures와 원자간 결합 패턴을 추론
  - 또한 스펙트럼 데이터베이스에서 구조적으로 유사한 분자를 검색하여 global contextual clues를 활용

- IR 스펙트럼은 다양한 화학적 정보와 함께 제공되는 경우가 많음
  - 예: 원자 종류 / 분자 골격(scaffold) / 탄소 개수 등
  - 그러나 기존 방법들 (Transformer 기반)은 이러한 정보를 유연하게 통합하는 데 한계가 있으며, 이를 반영하기 위해서는 모델 구조의 재설계, 재학습, 재검증이 필요함

→ LLM 기반 Agent 프레임워크의 필요성

1. 인간 전문가의 통합적 분석 워크플로우를 모사하고
2. 다양한 화학적 정보를 유연하게 통합할 수 있어야 한다.

Problem Setup

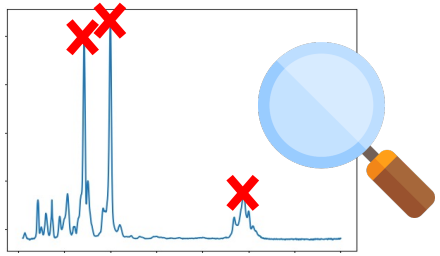
- 가정: 미지 물질의 **IR 스펙트럼만을 사용하여** SMILES를 예측
  - 실제 화학식(ground-truth chemical formula)은 주어지지 않음

- **왜 화학식을 가정하지 않는가?**
  - 일부 연구는 IR 스펙트럼과 함께 정확한 화학식이 항상 제공된다고 가정하지만,

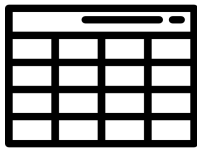
→ **실제 상황에서는 미지 물질의 정확한 화학식을 얻기 어렵다**

- 화학식을 얻기위해 질량분석(MS)이 일반적으로 사용되지만
  - 비용과 시간이 많이 소요됨
  - 해석이 복잡함
  - 정확한 화학식을 도출하는 것 역시 여전히 어려운 문제임

Preliminary

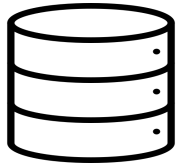


- **Tools**
  - **IR Peak Table Assigner**: 스펙트럼에서 피크를 추출하고 IR absorption table을 기반으로 관련된 부분구조를 매핑
  - **IR Spectra Retriever**: 입력 스펙트럼과 유사한 IR 스펙트럼을 IR 스펙트럼 DB에서 검색



Wavenumber (cm^{-1})	Substructure
3700–3584	alcohol (O–H)
3550–3200	alcohol (O–H)
3500–3400	primary amine (N–H)
3400–3300	aliphatic primary amine (N–H)
3330–3250	aliphatic primary amine (N–H)
3350–3310	secondary amine (N–H)
3100–2900	carboxylic acid (O–H)
3200–2700	alcohol (O–H)
3000–2800	amine salt (N–H)

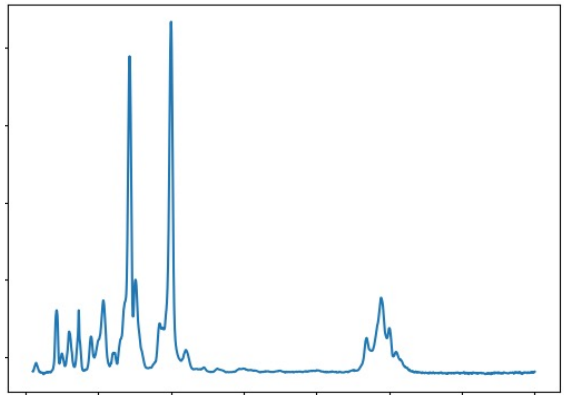
- **External Knowledge**
  - **IR Absorption Table:** 다양한 분자 작용기/부분구조에 대응하는 특성 흡수 파수 범위를 정리한 표
  - **IR 스펙트럼 DB:** 다양한 IR 스펙트럼과 이에 대응하는 분자의 SMILES 정보를 포함



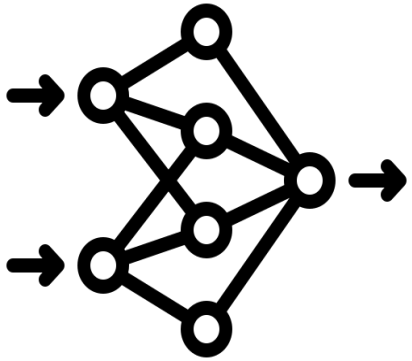
(IR Spectrum, SMILES)

Preliminary

- IR Spectra Translator



IR Spectrum $\mathcal{X}$



IR Spectra Translator

$\mathcal{C}$

: a set of SMILES candidates

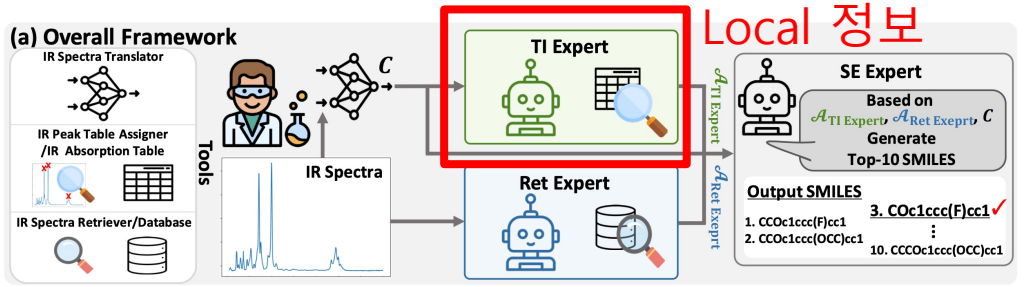
수천 개의 실수값(real-valued) IR absorption 데이터로부터 LLM이 직접 신뢰할 수 있는 SMILES를 도출하는 것은 어렵기 때문에,

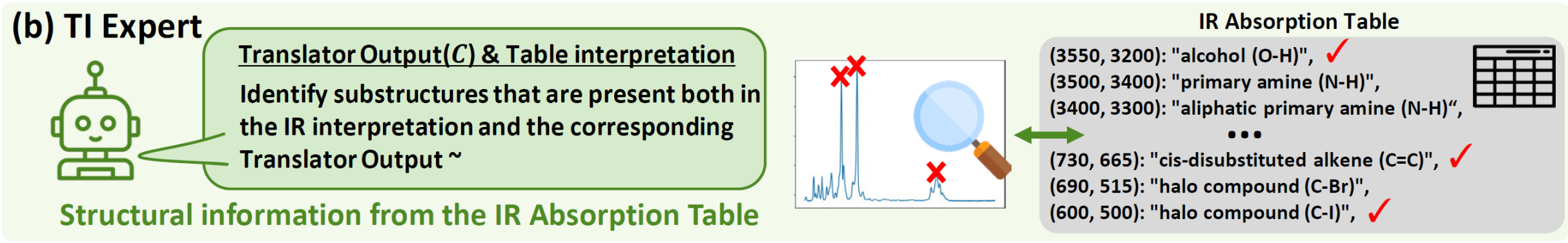
→ IR Spectra Translator를 활용하여 먼저 타당한 초기 구조 후보를 생성하고, 이후 이를 확장(expand)하고 정제(refine)

$$\mathcal{C} = \{\mathbf{s}_1, \ldots, \mathbf{s_K}\} = \mathrm{Transformer}(\mathcal{X})$$

Train 데이터로 pre-train

Table Interpretation (TI) Expert





- IR absorption table은 세밀한 Local 구조 정보를 제공하며, structure elucidation에 핵심적인 역할
- 이 표를 효과적으로 활용하려면 Peak를 정확히 식별해야 하며, 이는 LLM에게 도전적인 문제
- 따라서, IR Peak Table Assigner 도구를 활용하여 스펙트럼에서 Peak를 탐지하고, IR absorption table을 기반으로 해당 wavenumber 범위에 대응하는 부분구조를 할당

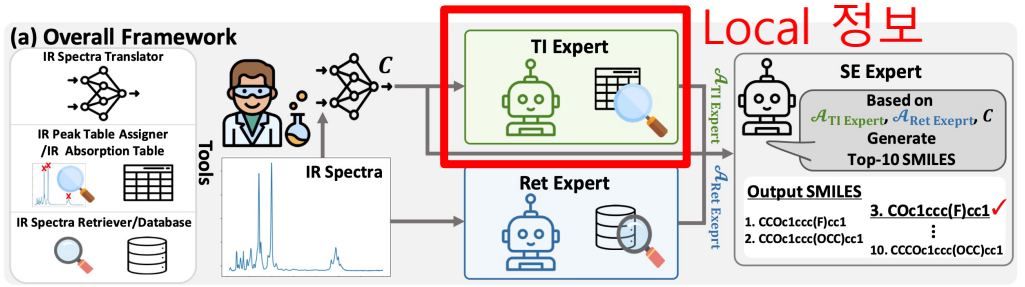
a set of SMILES candidates

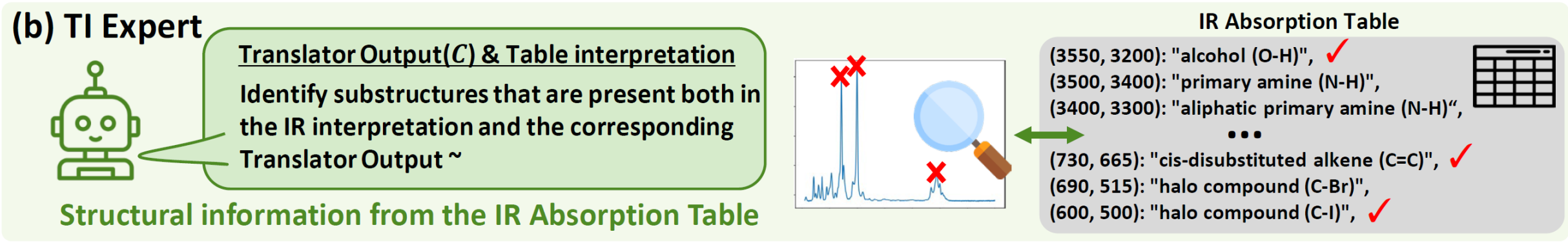
$$\mathcal{A}_{\text{TI Expert}} = \text{TI Expert}\left(\mathbf{P}_{\text{TI Expert}}, \text{IR Peak Table Assigner}(\mathcal{X}, \mathbf{T}), \mathcal{C}\right)$$

Task-specific prompt

IR Absoprtion Table

Table Interpretation (TI) Expert



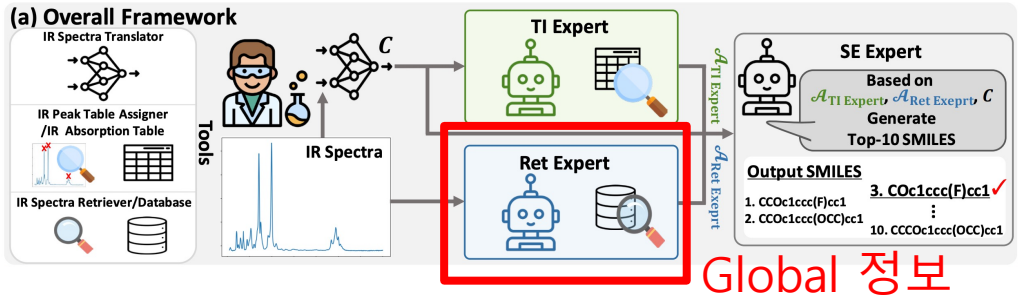


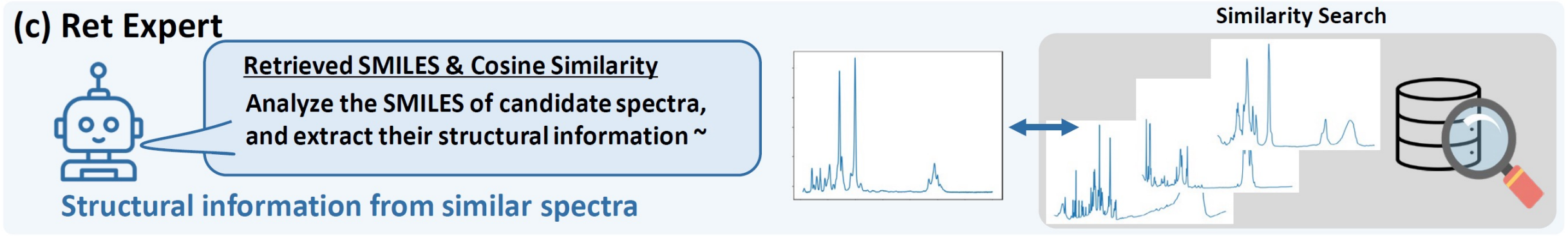
- But, IR Absorption table은 유용하지만, 표 기반 해석에는 본질적인 한계가 존재
  - IR 스펙트럼에는 noise가 포함될 수 있으며, 여러 부분구조가 유사한 wavenumber 영역을 공유하기 때문에 peak 할당 과정에서 모호성이 발생함
- 이러한 오해석 가능성을 완화하기 위해, IR Peak Table Assigner의 출력과 SMILES 후보를 비교하고, 공통 부분구조를 식별하며, 각 매칭에 대해 간단한 근거와 함께 신뢰도 점수를 산출하도록 유도 (*e.g., substructure → confidence → brief rationale)*

→ **표 기반 해석의 신뢰성을 향상**

- For Fc1ccc(Cl)c(Br)c1
  - C-F → High confidence → The "Fc" fragment indicates a fluorine directly bonded to the aromatic ring, matching the **1200–1000 cm^{-1} fluoro absorption.**
  - C-Cl → High confidence → The presence of a **chlorine substituent on the ring is consistent with the halo (C–Cl) absorption at 760–540 cm^{-1}.**

Retriever (Ret) Expert

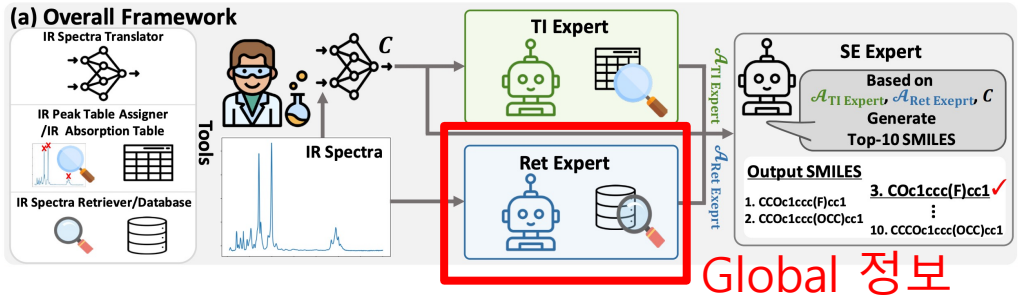


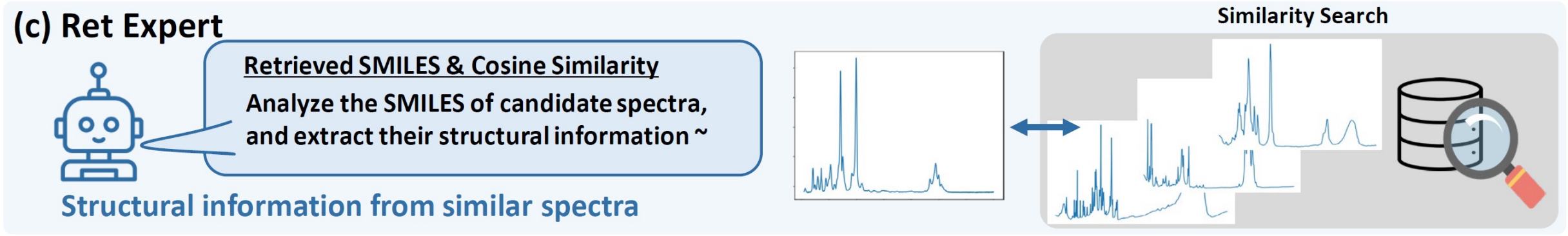


- TI Expert로부터 얻는 Local 구조 정보만으로는 완전한 분자 구조를 규명하기에 부족
- Ret Expert는 데이터베이스에서 유사한 IR 스펙트럼을 갖는 분자 구조들을 활용하여 global structural context을 제공 → Local substructure를 보다 완전한 분자 구조로 연결
  - (1) 먼저 타겟 스펙트럼과 데이터베이스 내 모든 스펙트럼 간의 코사인 유사도를 계산
  - (2) 이후 가장 유사한 상위 N개의 스펙트럼을 해당 SMILES 구조와 함께 검색

$$\{\text{candi}_1 : \text{sim}_1, \ldots, \text{candi}_N : \text{sim}_N\} = \text{IR Spectra Retriever}(\mathcal{X})$$

Retriever (Ret) Expert





$$\mathcal{A}_{\text{Ret Expert}} = \text{Ret Expert}\,(\mathbf{P}_{\text{Ret Expert}}, \text{IR Spectra Retriever}(\mathcal{X}))$$

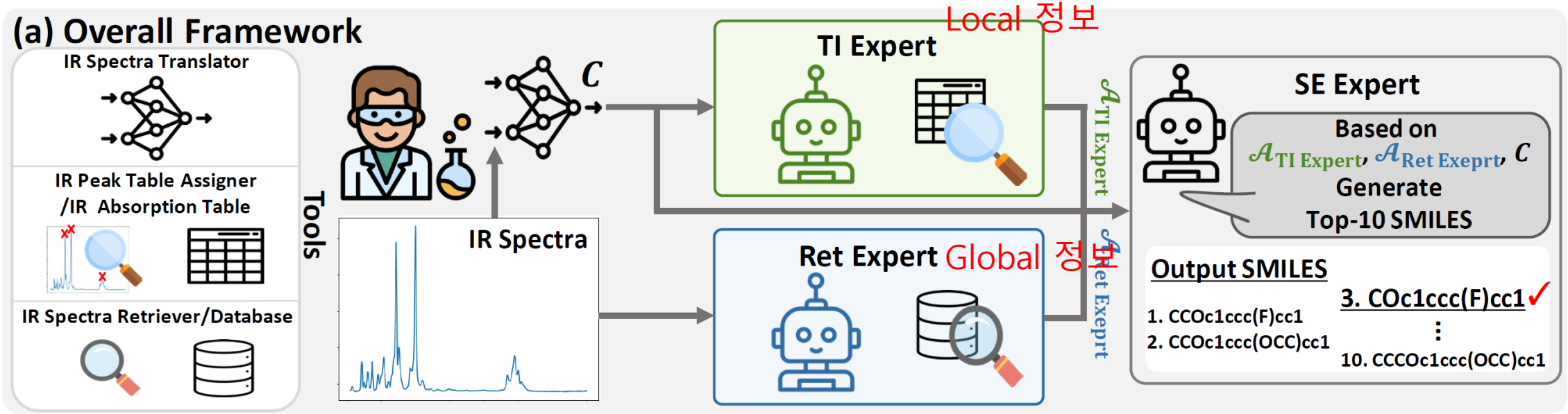
Task-specific prompt

- Ret Expert의 출력: 상위 N개로 검색된 SMILES들 사이에서 공통적으로 나타나는 구조적 특징 포함

→ 분자 구조 추론을 위한 global contextual clues 제공

All candidate SMILES share an aromatic core bearing **a trifluoromethyl group (–C(F)(F)F)**, suggesting that the target compound likely features **a benzene ring substituted with a CF3 group.** ... This pattern indicates that the target spectrum's molecule is an aromatic system with electron-donating and/or electron-withdrawing substituents that can influence IR absorptions (e.g., CF3 stretches near 1200 cm^{-1}, possible N–H stretches for amines, and C≡N stretches). The similarities in the candidates' SMILES imply that the target likely shares these structural motifs with variations in the nature and position of these substituents.

Structure Elucidation (SE) Expert



$$\mathcal{A}_{\text{SE Expert}} = \text{SE Expert}\left(\mathbf{P}_{\text{SE Expert}}, \mathcal{A}_{\text{TI Expert}}, \mathcal{A}_{\text{Ret Expert}}, \mathcal{C}\right)$$

Task-specific prompt | Output of TI Expert | Output of Ret Expert | SMILES candidates from Translator

- SE Expert는 TI Expert와 Ret Expert의 출력을 모두 기반으로 통합적인 구조 추론을 수행
  - Local 분자 구조와 Global 분자 구조를 모두 통합하는 포괄적인 추론 과정

→ 최종적으로 상위 K개의 예측된 분자 구조에 대한 순위 리스트를 생성

Experiments: Dataset & Evaluation Protocol

- 9,052 experimental IR spectra from NIST (National Institute of Standards and Technology) IR Spectral Library
  - 고체/액체/기체 모두 포함 / stereochemistry, ionic compound, mixture도 제외하지 않음
    → 실제 실험 환경과 유사한 설정을 유지
- **Knowledge Base:** Training set
- **Backbone LLM**: GPT-4o-mini, GPT-4o, o3-mini
- **Random Split**: train/valid/test of 80/10/10%
- SMILES를 InChI 표현으로 변환한 후, Top-K 정확도(K=1, 3, 5, 10)를 계산함 (정답 SMILES가 Top-K에 포함되어있는지)
  - SMILES는 같은 분자라도 **여러 가지 문자열 표현이 가능**

Experiments

*Single: 하나의 LLM이 모든 expert 역할을 하도록

Method	Agent	Top-K Accuracy			
		Top-1	Top-3	Top-5	Top-10
Transformer	-	0.098 (0.007)	0.169 (0.000)	0.176 (0.003)	0.176 (0.003)
IR-Agent (GPT-4o-mini)	single	0.072 (0.008)	0.118 (0.002)	0.133 (0.002)	0.157 (0.003)
	multi	0.093 (0.005)	0.152 (0.003)	0.167 (0.005)	0.176 (0.005)
IR-Agent (GPT-4o)	single	0.083 (0.004)	0.135 (0.002)	0.165 (0.007)	0.194 (0.008)
	multi	0.093 (0.007)	0.153 (0.005)	0.177 (0.005)	0.204 (0.005)
IR-Agent (o3-mini)	single	0.087 (0.006)	0.153 (0.005)	0.179 (0.002)	0.197 (0.004)
	multi	**0.103** (0.005)	**0.178** (0.007)	**0.199** (0.004)	**0.216** (0.001)

- Transformer 대비 성능 향상
- Multi-agent 프레임워크는 Single-agent 방식보다 일관되게 우수한 성능을 보임

Chemical Information	o3-mini				IR-Agent (single) (o3-mini)				IR-Agent (multi) (o3-mini)			
	Top-1	Top-3	Top-5	Top-10	Top-1	Top-3	Top-5	Top-10	Top-1	Top-3	Top-5	Top-10
No Knowledge	0.073 (0.010)	0.131 (0.011)	0.157 (0.011)	0.185 (0.005)	0.087 (0.010)	0.153 (0.011)	0.179 (0.011)	0.197 (0.005)	0.103 (0.005)	0.178 (0.007)	0.199 (0.004)	0.216 (0.001)
Scaffold	0.096 (0.002)	0.160 (0.006)	0.177 (0.003)	0.198 (0.003)	0.112 (0.003)	0.195 (0.009)	0.208 (0.008)	0.228 (0.010)	0.118 (0.003)	0.208 (0.009)	0.232 (0.008)	0.258 (0.010)
Carbon Count	**0.105** (0.009)	0.158 (0.014)	0.186 (0.014)	0.214 (0.013)	0.121 (0.003)	0.177 (0.008)	0.194 (0.010)	0.219 (0.009)	0.123 (0.003)	0.190 (0.0005)	0.215 (0.009)	0.252 (0.007)
Atom Types	0.104 (0.011	**0.182** (0.007)	**0.209** (0.005)	**0.237** (0.003)	**0.123** (0.006)	**0.208** (0.003)	**0.235** (0.011)	**0.266** (0.009)	**0.127** (0.006)	**0.213** (0.003)	**0.250** (0.011)	**0.278** (0.009)

- 화학적 정보는 agent들의 프롬프트에 **텍스트 형태로 추가**
- 간단한 텍스트 형태의 화학 정보만 추가하더라도 (모델 구조 변경이나 재학습 없이도) 분자 구조를 보다 정확하게 예측 가능

→ 이는 다양한 형태의 화학 정보를 유연하게 통합할 수 있는 agent 프레임워크임의 장점 부각

Experiments

Translator Output(C)

'Nc1ccc(C(F)(F)F)cc1Br', 'Nc1ccc(Br)c(C(F)(F)F)c1', 'Nc1ccc(N)c(C(F)(F)F)c1'

TI Expert

• C-F group → High confidence → All three SMILES include a trifluoromethyl (CF_3) substituent, which corresponds to the **IR peaks between 1200–1000 cm^{-1} typically seen for C–F bonds.**
• Halogen substitution (Br) → Low confidence → SMILES 1 and 2 show bromine on the aromatic ring, but **the IR interpretation specifies C–Cl**; while both are halogens, the match is imprecise.

Ret Expert

All candidate SMILES share an aromatic core bearing **a trifluoromethyl group (–C(F)(F)F)**, suggesting that the target compound likely features **a benzene ring substituted with a CF3 group.** ... This pattern indicates that the target spectrum's molecule is an aromatic system with electron-donating and/or electron-withdrawing substituents that can influence IR absorptions (e.g., CF3 stretches near 1200 cm^{-1}, possible N–H stretches for amines, and C≡N stretches). The similarities in the candidates' SMILES imply that the target likely shares these structural motifs with variations in the nature and position of these substituents.

SE Expert

1. Nc1ccc(C(F)(F)F)cc1Br, 2. Nc1ccc(Br)c(C(F)(F)F)c1 ..., 4. **Nc1ccc(Cl)c(C(F)(F)F)c1** ✓, ..., 10. Nc1ccc(O)c(C(F)(F)F)c1

- TI Expert는 Translator 출력과 IR table 해석이 모두 C–F 결합을 일관되게 가리키기 때문에 C–F 그룹을 높은 신뢰도로 판단
- 반면, Br과 같은 할로겐 치환의 경우 IR table 해석에서는 C–Cl 결합이 제시되어 증거가 상충하므로 낮은 신뢰도로 판단
- Ret Expert는 검색된 후보들 사이에서 CF_3가 치환된 벤젠 고리를 지배적인 구조적 모티프로 식별함으로써 global 맥락을 제공

Periodic Complex Stochastic Processes for Retrieving Atomic Structures of Unknown Matters

Under Review

Gyoung S. Na, Chanyoung Park

왜 분광 기반 구조 추론은 어려운가?

- **간접 관측 문제 (Inverse Problem)**
  - 구조를 직접 보는 것이 아님
  - 신호 → 구조로 역추론해야 함
- **구조적 모호성 (Ambiguity)**
  - 서로 다른 구조가 비슷한 스펙트럼을 만들 수 있음
  - 단일 peak는 구조를 결정하지 못함
  - 전체 패턴을 종합적으로 해석해야 함
- **신호의 집합성 (Aggregation)**
  - Spectrum은 여러 진동 모드 신호의 합
  - 환경 및 실험 조건에 따른 변화 포함
  - Noise 및 측정 오차 존재

그렇다면 우리는 분광 데이터를 AI로 어떻게 모델링해야 할까?

👉 **강한 Inductive bias가 필요**

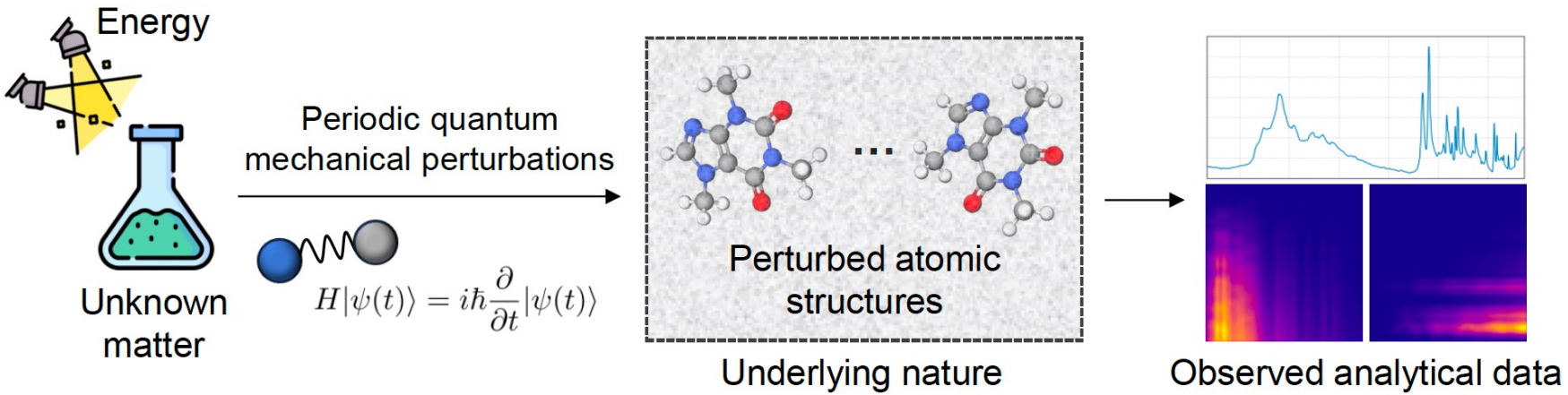
1. 전문가의 해석 과정을 모델링 하자 → IR-Agent (ICLR 2026)
   - Spectrum은 전문가가 읽는 신호 (local + global pattern)
   - Reasoning 과정을 학습하자

2. Spectrum의 생성 메커니즘을 모델링 하자 → PCSP (Under Review)
   - 여러 진동 모드의 신호가 중첩된 결과
   - 스펙트럼의 생성 과정(generative process)을 모델링하자

PCSP - Motivation

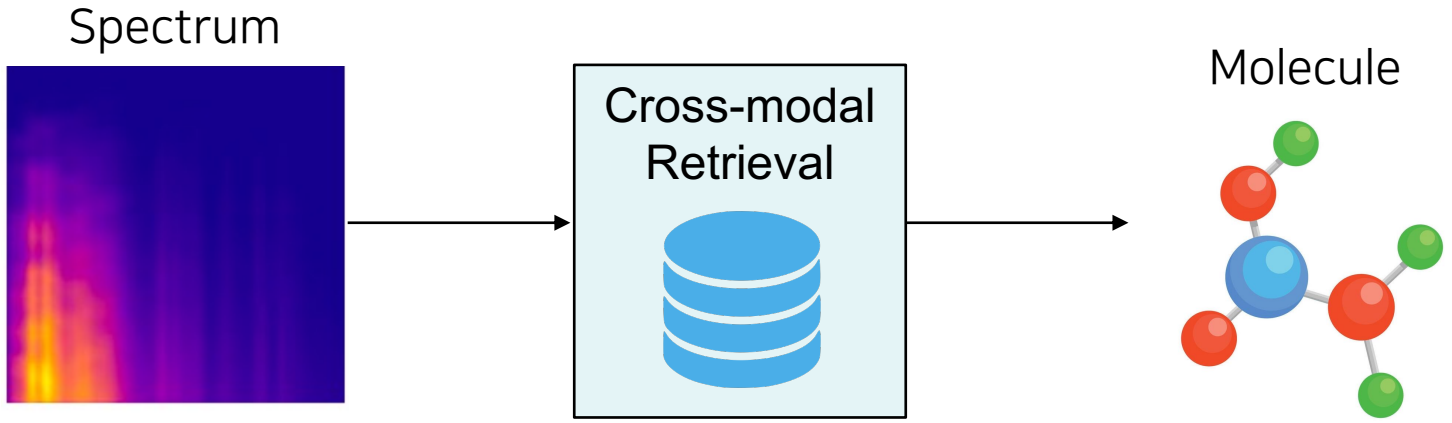
- Spectrum은 단일 equilibrium state의 결과가 아니다
- 실제로는 **"주기적으로 perturbation된 여러 구조들의 집합"에서 나온다**

→ **즉, 관측 데이터는 perturbed state들의 집합** (여러 순간의상태가 섞여서 최종 스펙트럼이 만들어짐)



- **기존연구**: Perturbed states를 무시하고 equilibrium state만 학습
- **본 연구**: Atomic structure embedding을 periodic complex stochastic process로 확장

문제 정의



- **Task**: 주어진 Spectrum 데이터에 가장 맞는 Molecule 구조 찾기

Retrieval Task	Dataset	Input Query	Target Data	# of Samples	Data Source
Sequence-to-graph information retrieval	QMe14S-IR	IR spectrum	Molecule	186,102	DFT calculation
	QMe14S-RM	Raman spectrum	Molecule	186,102	DFT calculation
	NIST	IR spectrum	Molecule	8,832	Experiment
	OpenXRD	X-ray diffraction patterns	Crystalline material	872	Experiment
Image-to-graph information retrieval	2DNMRGym	NMR scattering image	Molecule	11,825	Experiment
	INS-MAT	Neutron scattering image	Crystalline material	9,928	DFT calculation

Overview

- 양자역학 적으로 진동은 **주기적**이기 때문에, 주기적으로 변하는 stochastic process를 만들자

$$p(\mathbf{x}_t|\mathbf{x}_{t-1},\ldots,\mathbf{x}_0) = p(\mathbf{x}_{t+L}|\mathbf{x}_{t+L-1},\ldots,\mathbf{x}_0) \quad (L\text{: 주기})$$

- 이를 복소수 공간에서 정의 → Periodic Complex Stochastic Process (PCSP)
  - 이전 구조를 복소평면에서 조금 회전시키고, 거기에 적절한 확률적 진동을 더한 모델

$$p(\mathbf{s}_t) = \mathcal{CN}(\mathbf{s}_t; \alpha_t \mathbf{s}_{t-1}, \gamma^2(1-\alpha_t^2)\mathbf{I})$$

회전에 다양성을 주기위해 Noise 추가

t 시점의 perturbed structure embedding (복소수 벡터)

이전 상태를 복소수 계수 α_t로 회전

생성식

$$s_t = \alpha_t s_{t-1} + \gamma\sqrt{(1-\alpha_t^2)}\epsilon_t$$

$\alpha_t = e^{i\delta} = \cos\delta + i\sin\delta$ (복소평면에서 회전연산 – 오일러 공식) $\delta = 2\pi/L$

L step마다 한바퀴 도는 구조

- Cross-modal Retrieval

Algorithm 2 Information Retrieval Process of CVCR

Input: analytical data $\mathbf{x}$,
periodicity criterion $\beta \in \{0.9, 0.95, 0.99\}$.
Output: index of the retrieved atomic structure j^*.
for $\mathcal{A} \in \mathcal{D}_{\mathcal{A}}$ **do**
  $\tilde{\mathbf{z}}_0, \tilde{\mathbf{z}}_1, \ldots, \tilde{\mathbf{z}}_{L-1} = \texttt{Normal-PCSP}(h_{\mathcal{A}}(\mathcal{A}), \beta)$ ← L개의 perturbed embedding 생성
  $\mathbf{z} = \sum_{t=0}^{L-1} \alpha_t \tilde{\mathbf{z}}_t$
  $\mathcal{S} = \mathcal{S} + [h_{\mathcal{Q}}(\mathbf{x}) \cdot \mathbf{z}]$ // Append the calculated similarity.
end for
Return: argmax $\mathcal{S}$

Future work: Multimodal Spectroscopic Data Modeling

- 기존 연구는 대부분 IR / NMR / MS 중 한개 만 사용 (single modality 기반 모델)
- 하지만 실제 화학자들은 여러 spectroscopy 정보를 동시에 사용해서 구조를 추론
  - IR → functional group
  - NMR → 구조 skeleton
  - MS → molecular formula

Overview of the data available for the different modalities.

Modality	Subtype	Data Description
IR	Spectrum	Vector size 1800: from 400cm^{-1} to 4000cm^{-1} with a 2cm^{-1} resolution
^{1}H-NMR	Spectrum	Vector size 10'000: from 10ppm to -2pmm with a 0.0012ppm resolution
	Annotated Spectrum	Start, End, Centroid, Integration and Type of each peak
^{13}C-NMR	Spectrum	Vector size 10'000: from 230ppm to -20pmm with a 0.025ppm resolution
	Annotated Spectrum	Centroid and Intensity of each peak
HSQC-NMR	Spectrum	Matrix shape 512×512: ^{1}H-NMR-dim: Vector size 512: from 10ppm to -2pmm with a 0.023ppm resolution ^{13}C-NMR-dim: Vector size 512: from 230ppm to -20pmm with a 0.488ppm resolution
	Annotated Spectrum	X, Y coordinates and integration of each peak
Positive MS/MS	Spectrum	m/z & Intensity of each peak
	m/z Annotations	Chemical formula corresponding to the m/z of each peak
Negative MS/MS	Spectrum	m/z & Intensity of each peak
	m/z Annotations	Chemical formula corresponding to the m/z of each peak

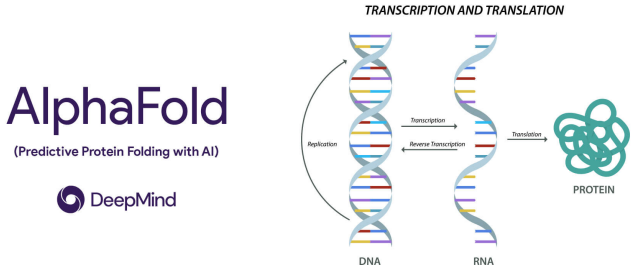
(NeurIPS 2024 D&B Track) Unraveling Molecular Structure: A Multimodal Spectroscopic Dataset for Chemistry

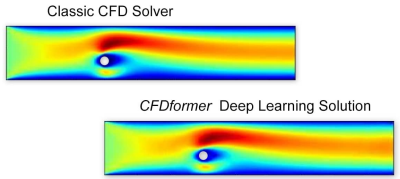
Outline

- 소재 합성 모델링
- 분광 데이터 분석
- Agentic AI for Science
- Fragmentation 기반 Molecule Representation Learning

과학의 목표: 메커니즘 발견

- 과학은 관측 데이터로부터 지배 방정식(Governing Equation)을 발견하는 과정 (예. F=ma)
- 지배 방정식은 시스템의 본질을 **해석 가능하고(explainable) 일반화 가능하게(extrapolatable)** 표현
- **현대 AI의 성과와 한계**
  - AlphaFold, 분자 설계, 유체 시뮬레이션 등에서 뛰어난 성과
  - 그러나 딥러닝 기반 AI는 본질적으로 $y = f_\theta(x)$
  - 내부 메커니즘 해석 어려움, OOD 일반화 취약

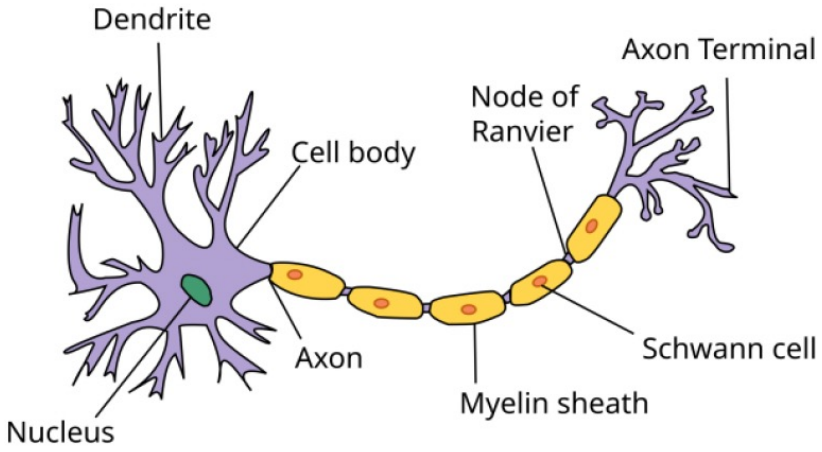




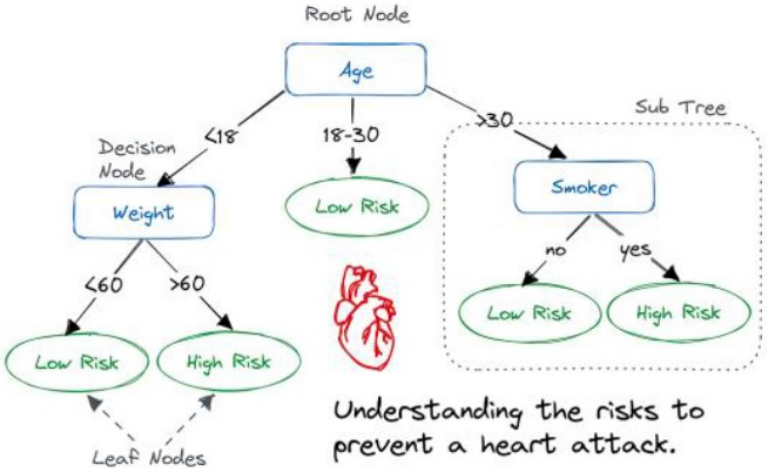
→ 우리는 "정확한 예측 모델"이 아니라 "과학 법칙을 스스로 발견하는 AI"가 필요

Connectionism vs. Symbolism

- Connectionism
  - 생물학적 뉴런 모사
  - 데이터로부터 패턴 학습
  - 함수 근사 중심
  - 결과는 비해석적 블랙박스



- Symbolism
  - 규칙과 논리 구성
  - 명시적 수식 표현
  - 추론 과정이 설명 가능
  - 지배 방정식 발견과 직접 연결



Connectionism의 표현력 + Symbolism의 해석 가능성
→ **설명 가능하고 Extrapolation 가능한 지배 방정식 발견**

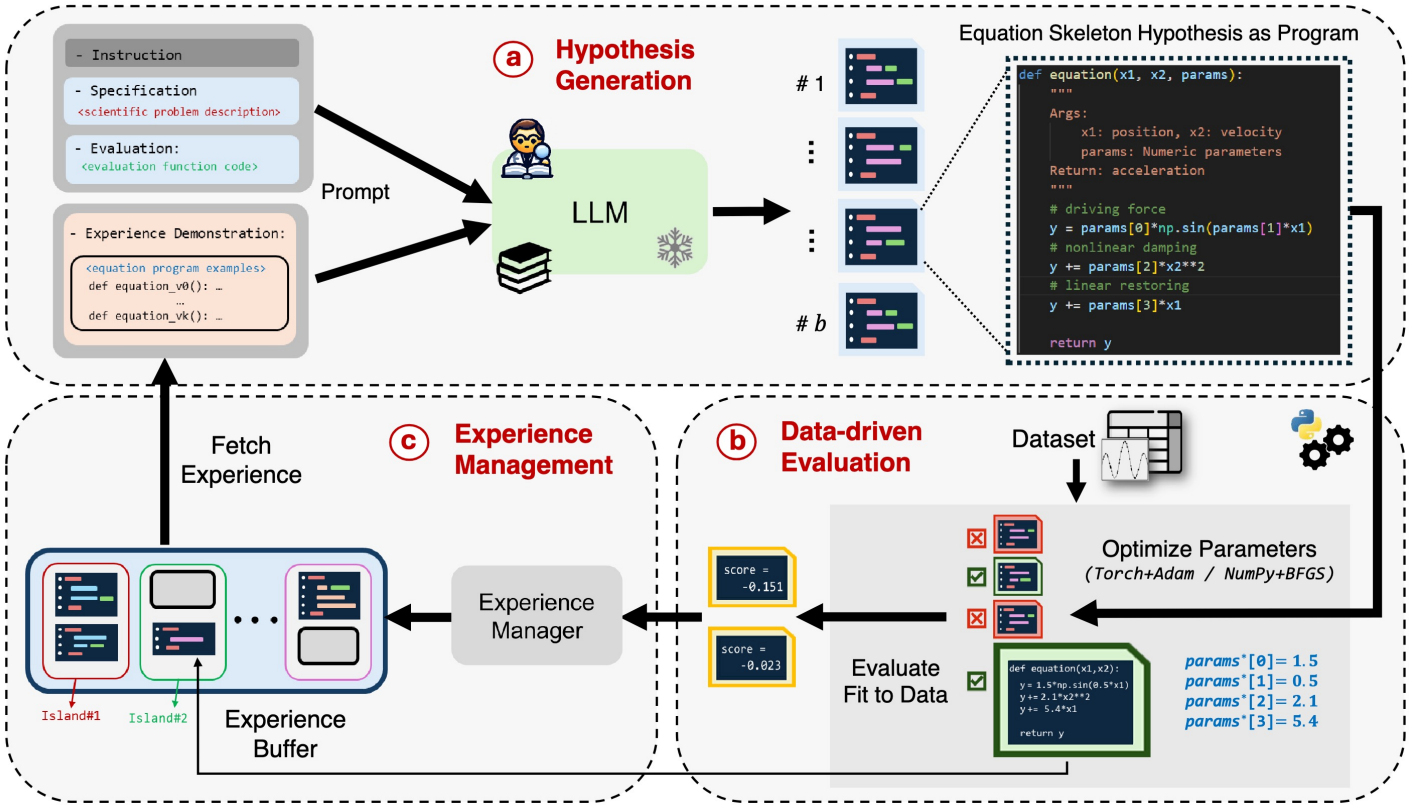
(ICLR 2025 Oral) LLM-SR: Scientific Equation Discovery via Programming with Large Language Models

- LLM을 활용해 과학 방정식을 '프로그램 형태'로 생성하고, 데이터 기반 최적화 + Evolutionary search를 결합해 새로운 수식을 발견하는 방법론
- 문제 설정
  - Given: (x_i, y_i)
  - Goal: 그 관계를 설명하는 해석가능한 수식 $\tilde{f}(x)$를 찾는것
  - 즉, 단순히 예측 모델을 만드는 게 아니라, 물리적 의미를 가지는 방정식 자체를 발견하는 문제
- 예시: Material Stress Behavior (재료 응력-변형률 관계)
  - Given: $x =$ (변형률 (ϵ), 온도 (T)), $y =$ 응력 (σ)
  - Output: $\tilde{f}$: (ϵ,T)→σ
  - 발견된 수식 예시

$$\sigma = E_0 e^{-\alpha T} \epsilon + \beta \epsilon^3$$

(온도 의존 비선형 응력–변형률(stress–strain) 모델)

LLM-SR: Overview



- Step 1: Hypothesis Generation

```
def f(x, params):
    return params[0] * torch.sin(x) + params[1] * x**3
```

  - 수식의 구조만 생성 (계수는 placeholder)
  - 이전 잘된 수식도 in-context로 제공

- Step 2: Data-driven Evaluation

$$\text{Score} = MSE(y - \hat{y})$$

  - 수식의 구조는 LLM이, 숫자는 optimizer (torch+Adam)가 찾음

- Step 3: Experience Buffer
  - 좋은 수식들을 buffer에 저장 (“수식:점수” 포맷)
  - 다음 prompt에 in-context example로 제공
  - LLM이 Mutation + Cross-over

LLM-SR의 한계점

- LLM-SR: 하나의 LLM이 1) 프로그램 생성, 2) 수식 조합, 3)계수 최적화
- 문제
  - 1) Program(text) 기반 탐색의 한계: LLM-SR은 수식을 text program으로 생성
    - search space 정의 불명확
    - systematic symbolic search 불가능
    - sampling 기반 탐색
  - 2) LLM prior 의존
    - LLM이 익숙한 수식 형태는 잘 찾음
    - LLM이 모르는 시스템에서는 탐색 어려움

    → 사실상 LLM이 알고 있는 수식 템플릿 조합
  - 3) 구조 발견 대신 근사에 머무름
    - 단순히 prediction error 최소화
    - 실제 governing equation 발견 실패 가능

    → OOD extrapolation 취약

Machine Collective Intelligence (MCI)

- 1) Abstract Syntax Tree (AST) 기반 수식 표현
  - 수식을 AST (Abstract Syntax Tree) 구조로 표현

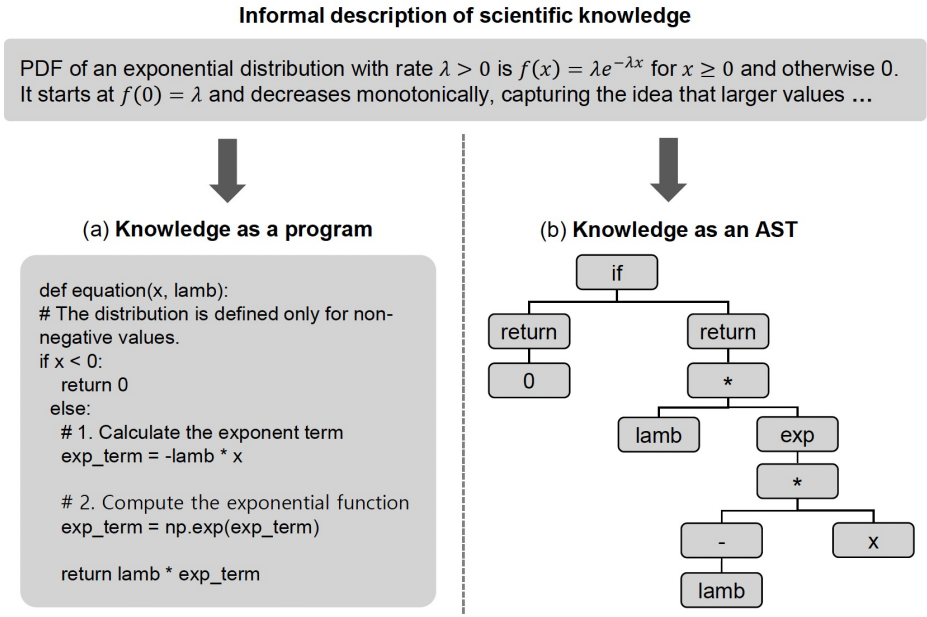
  → structured symbolic search 가능
- 2) Multi-Agent Collective Exploration
  - 단일 LLM이 아니라 여러 reasoning agent가 병렬 가설 생성
  - 서로 다른 초기 가설 → 탐색 다양성 증가 (특정 Prior 고정 x)
- 3) Only 오차 최소화 → 오차 및 구조 최소화 + 설명 가능성
  - AST (Abstract Syntax Tree) 기반 구조평가 (깊이: 복잡도)

$$s(AST_k, \mathcal{D}) = -\left\{ \sum_{(\mathbf{x},y)\in\mathcal{D}} (y - AST_k(\mathbf{x}))^2 + d_k + M_k \right\}$$

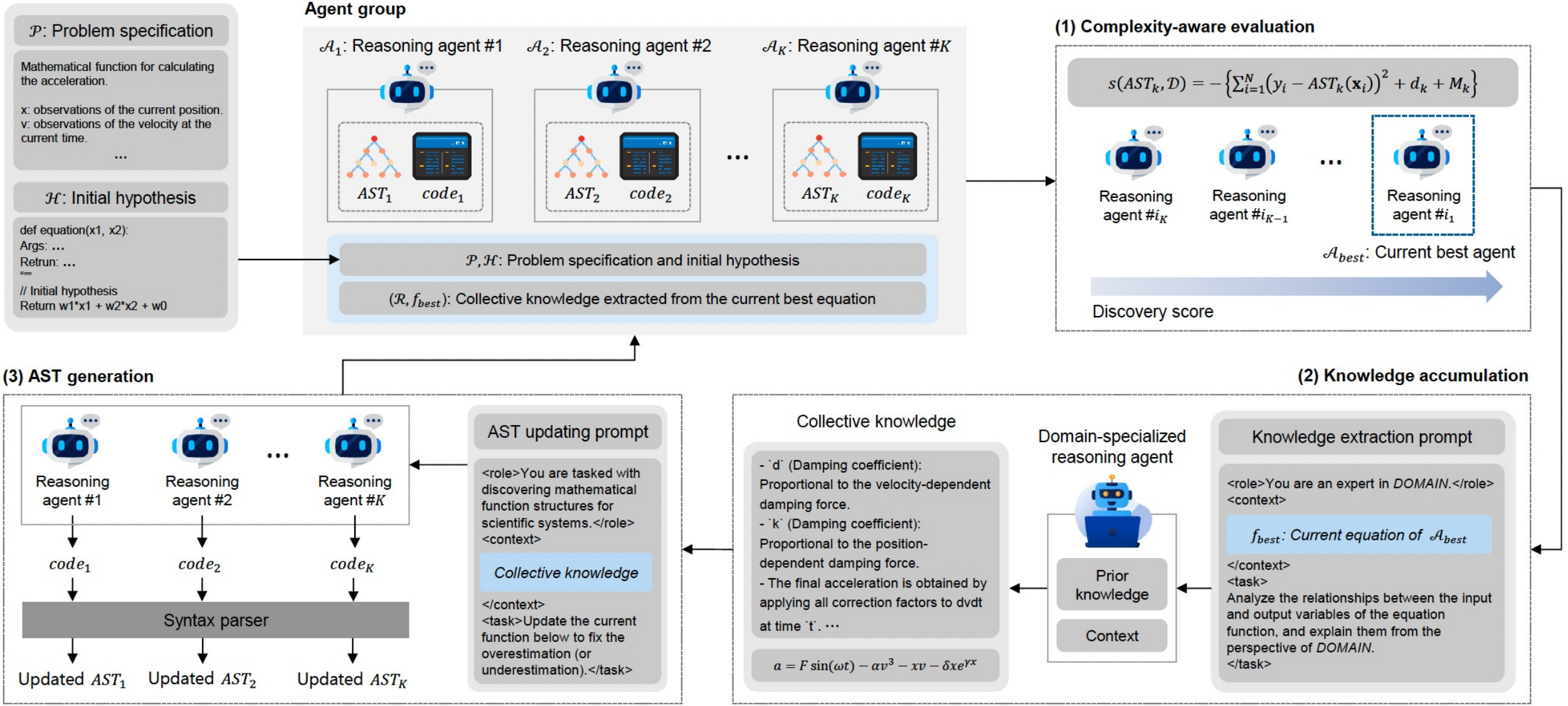
d:AST depth, M: Num. free parameters

  - best 수식 → 도메인 해석 → 자연어 지식으로 변환 (모든 agent가 공유)

  → prior 바깥 영역에서도 점진적 구조 학습 가능



MCI: Overview



Result

Domain	Problem Name (Target System)	Input Variables	Output Variable	Governing Equation
Physical science	**Chi2PDF** [34] (PDF of arbitrary chi-squared distributions)	• x: Data point • k: Degree of freedom	Probability density	$d_{x,k} = \frac{1}{2^{k/2}\Gamma(k/2)} x^{(k/2)-1} e^{-x/2}$
	NOSC [24] (Damped oscillator with nonlinear components)	• t: Time • x: Position • v: Velocity	Acceleration	$a = 0.3\sin(t) - 0.5v^3 - xv - 5xe^{0.5x}$
	NNN [1] (Nonlinear neural networks with sigmoid activation)	• x_1: 1st model input • x_2: 2nd model input	Model output	Fully-connected neural network with two hidden layers in Eq. (6)
	MSB [35] (Strain-stress relation in a metallic material)	• T: Temperature • ϵ: Strain	Mechanical stress	Unknown
Chemical science	**FHST** [36] (Flory–Huggins equation for polymer mixing)	• ϕ: Volume fraction • N: Chain length • χ: Interaction param.	Total change of Gibbs free energy	$\Delta G_m = RT\left[\frac{\phi}{N}\ln\phi + \psi\ln\psi + \chi\phi\psi\right]$, where $R = 8.314, T = 300, \psi = 1-\phi$
	BDC [37] (Lithium-ion battery)	Six measurement parameters (cycle, voltage, etc.)	Discharge capacity	Unknown
	SFL [37] (Fatigue limit of alloy compositions)	Weight Percents of ten elements and tempering temperature	Fatigue limit	Unknown
	NOMC [38] (Chemical reactor for non-oxidative methane conversion)	Six engineering parameters (pressure, time, etc.)	C_2 conversion yield	Unknown
Biology	**ECBG** [24] (Differential equation for bacterial growth rate of Escherichia coli)	• b: Population density • s: Substrate conc. • T: Temperature • pH: pH level	Bacterial growth rate	Differential equation for bacterial growth rate in Eq. (9)
	HHM [39] (Differential equation for membrane potential in Hodgkin-Huxley model)	Five biophysical parameters (membrane potential, gating probability, etc.)	Rate of potential change	Differential equation for Hodgkin–Huxley membrane potential dynamics in Eq. (**??**)

Problem	GPlearn	PySR	LLM-SR	MSI	MCI
Chi2PDF	0.6629 (±0.0020)	0.4656 (±0.0184)	0.0275 (±0.0041)	0.6915 (±0.0838)	**2.23E-6 (±3.13E-6)**
NOSC	0.4159 (±0.0009)	0.1713 (±0.1741)	0.0276 (±0.0032)	0.0642 (±0.0062)	**1.01E-6 (±2.83E-7)**
NNN	0.5049 (±0.0356)	N/A	0.0048 (±0.0017)	0.0538 (±0.0070)	**0.0015 (±0.0002)**
MSB	0.1962 (±0.0206)	0.1613 (±0.0101)	0.0401 (±0.0045)	0.0791 (±0.0084)	**0.0281 (±0.0033)**
FHST	0.0820 (±0.0202)	0.0783 (±0.0091)	0.0117 (±0.0010)	0.0674 (±0.0066)	**1.25E-7 (±1.77E-7)**
BDC	0.0989 (±0.0044)	0.0149 (±0.0000)	0.0154 (±0.0038)	0.0137 (±0.0045)	**0.0083 (±0.0021)**
SFL	0.1844 (±0.0097)	0.1199 (±0.0047)	0.1131 (±0.0202)	0.0821 (±0.0076)	**0.0431 (±0.0012)**
NOMC	0.2638 (±0.0220)	0.2031 (±0.0001)	0.1823 (±0.0167)	0.1658 (±0.0147)	**0.0609 (±0.0052)**
ECBG	1.0087 (±0.0079)	1.4973 (±0.0001)	0.3182 (±0.0875)	0.3460 (±0.0151)	**0.0581 (±0.0211)**
HHM	0.5949 (±0.0048)	0.8869 (±0.0129)	2.24E-7 (±1.58E-7)	0.0004 (±0.0006)	**7.21E-8(±5.46E-8)**

- 전반적으로 큰 성능향상
- 특히, governing equation이 알려지지 않은 경우에 더 큰 성능향상

Chi2PDF

Ground-truth:

$$d_{x,k} = \frac{1}{2^{k/2}\Gamma(k/2)} x^{(k/2)-1} e^{-x/2}$$

LLM-SR:

$$d_{x,k} = \frac{1}{\left(\sqrt{|0.3298x|} + 0.3428|k| - 0.0758\right)\sqrt{2\pi}} e^{-g(x)/2}$$

MCI:

$$d_{x,k} = \frac{1}{2^{k/2}\Gamma(k/2)} \max(x, -0.0123)^{(k/2)-1}\, e^{-\max(x,-0.0123)/2}$$

NOSC

Ground-truth:

$$a = 0.3\sin(t) - 0.5v^3 - xv - 5xe^{0.5x}$$

LLM-SR:

$$a = \frac{0.1836\sin(t) - 0.1378v - 3.0572x - 1.4857x^2}{0.6086 - (-0.2001 + 0.0734\sin(t))v}$$

MCI:

$$a = \left(\frac{dv}{dt} - 0.0002xe^{-0.0004x}\right) e^{-0.0004x}$$

<찾아낸 수식>

Result

<점진적인 수식 update 과정>

$K = 1$

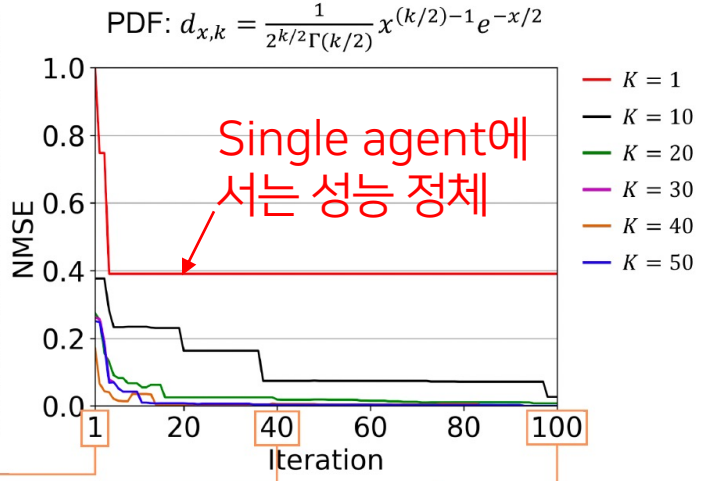
$$d_{x,k} = \frac{\text{ndtr}((x-\theta_3)/|\theta_1|)}{(1+(x-\theta_3)^2/(k+\theta_2))}$$

$K = 10$

$$d_{x,k} = \text{ndtr}\left(\frac{x-\theta_1}{\theta_2^2}\right)\frac{\theta_3}{\sqrt{k}} + \theta_4$$

$K = 50$

$$d_{x,k} = \frac{\sqrt{\theta_2 k/2\pi}e^{-(\theta_1 x-\theta_3)^2}}{2\theta_2\theta_4 e^{-\theta_2 k/2}}$$



$K = 1$

$$d_{x,k} = \text{cauchy}\left(\frac{x-\theta_3}{k+\theta_2}\right)(k+\theta_2)e^{|\theta_1|}\left|\frac{|\theta_1|}{(k+\theta_2)^2}\right|$$

$K = 10$

$$d_{x,k} = \frac{\text{ndtr}(\theta_3(x-\theta_5))}{(k(x-\theta_5)^2+\theta_3^2)^{1/2}}\theta_6 x^{2(k-\theta_7)}e^{-x/2}$$

$K = 50$

$$d_{x,k} = \frac{1}{2^{k/2}\Gamma(k/2)}x^{2((k/2)-1)}e^{-x^2/2}$$

$K = 1$

$$d_{x,k} = \text{cauchy}\left(\frac{x-\theta_3}{k+\theta_2}\right)(k+\theta_2)e^{|\theta_1|}\left|\frac{|\theta_1|}{(k+\theta_2)^2}\right|$$

$K = 10$

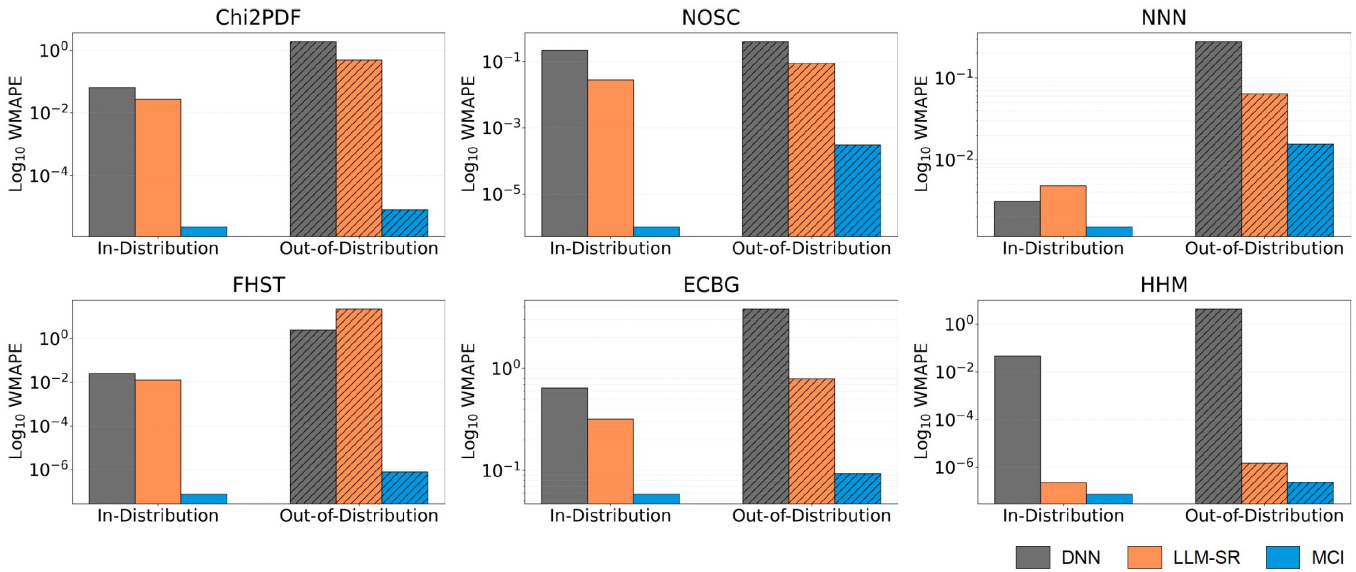
$$d_{x,k} = \frac{\theta_2^k}{\Gamma(k)}x^{k-1}e^{-x/\theta_1+\theta_3 x}$$

$K = 50$

$$d_{x,k} = \frac{1}{2^{k/2}\Gamma(k/2)}\max(x,\theta_1)^{(k/2)-1}e^{-\max(x,\theta_1)/2}$$

K가 커질수록 정답 equation의 핵심구조 정확히 복원

<Out-of-distribution>



- DNN은 특히 OOD에서는 일부 문제에서 크게 fail
- LLM-SR 역시 OOD에서 오차가 크게 증가

→ 기존 연구들은 Extrapolation에 실패

- 반면 MCI는 특히 OOD에서 기존 연구대비 낮은 오차

→ 구조적 법칙 발견 기반 접근의 강한 OOD 일반화성능 입증

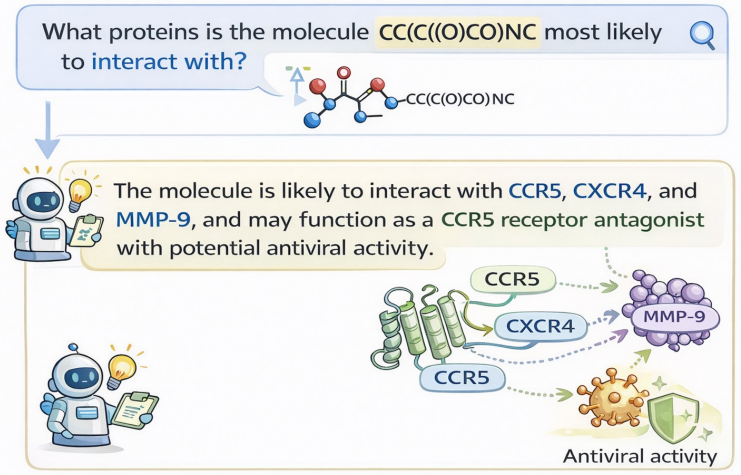
RAG-Enhanced Collaborative LLM Agents for Drug Discovery

AAAI 2026

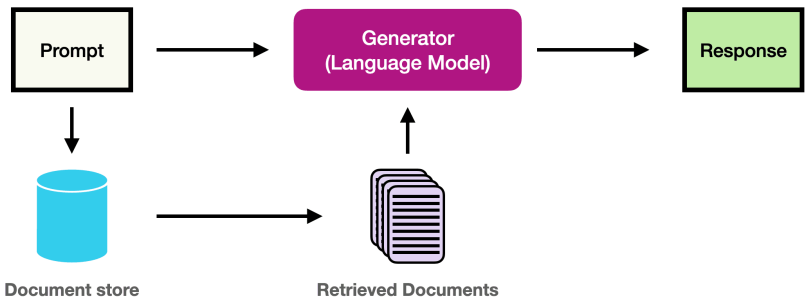
Namkyeong Lee, Edward De Brouwer, Ehsan Hajiramezanali, Tommaso Biancalani, Chanyoung Park, Gabriele Scalia

Introduction

- Task: General Molecular Question Answering (Molecular QA)
  - Drug-Target Prediction Task / Molecular Captioning Task / Drug Toxicity Prediction Task
- 기존 방법: Domain-specific fine-tuning
  - 장점: Task-specific performance 향상 / Molecule-specific reasoning 가능
  - 한계
    - 새로운 LLM 등장 시 반복적 fine-tuning 필요
    - 지속적으로 생성되는 실험/논문 지식 반영 어려움
    - General한 open-ended scientific reasoning에 취약
- Alternative: Retrieval-Augmented Generation (RAG)
  - RAG는 fine-tuning 없이 외부 지식을 inference 시점에 활용 가능
  - Challenge
    - Query molecule이 DB에 없을 가능성 높음
    - Multi-source 통합 필요 (외부 정보는 Multi-modality)
    - Real-world task들은 Open-ended reasoning 필요



<Drug-Target Prediction>



<RAG>

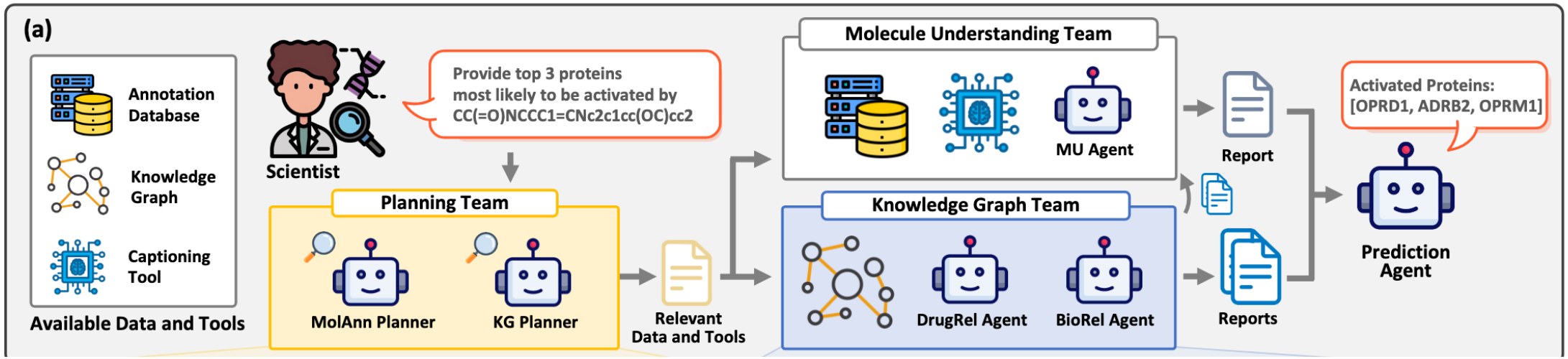
CLADD: Overview

- **문제 1: Fine-tuning 없이 drug discovery task를 수행할 수 있는가?** → zero-shot으로 해결
- **문제 2: Query molecule이 knowledge base에 없을 때 어떻게 할 것인가?**

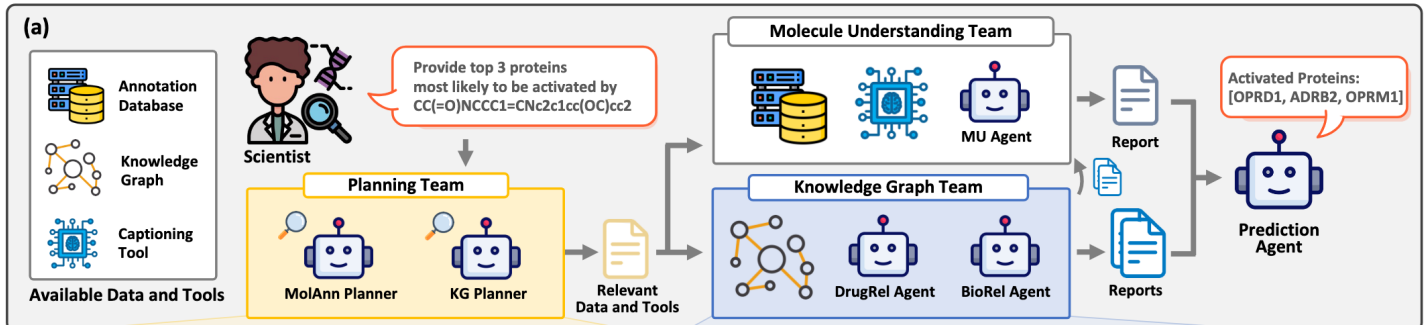
→ Anchor drug + structural similarity + 2-hop KG reasoning

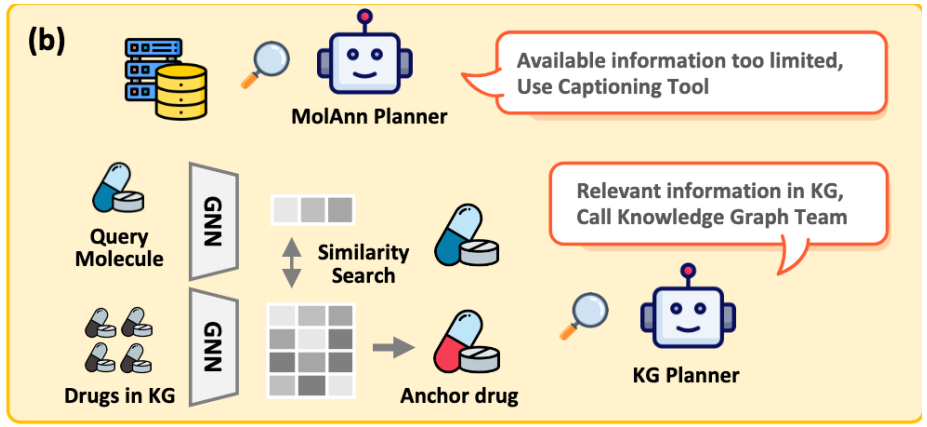
- **문제 3: Heterogeneous biochemical data를 어떻게 통합할 것인가?**

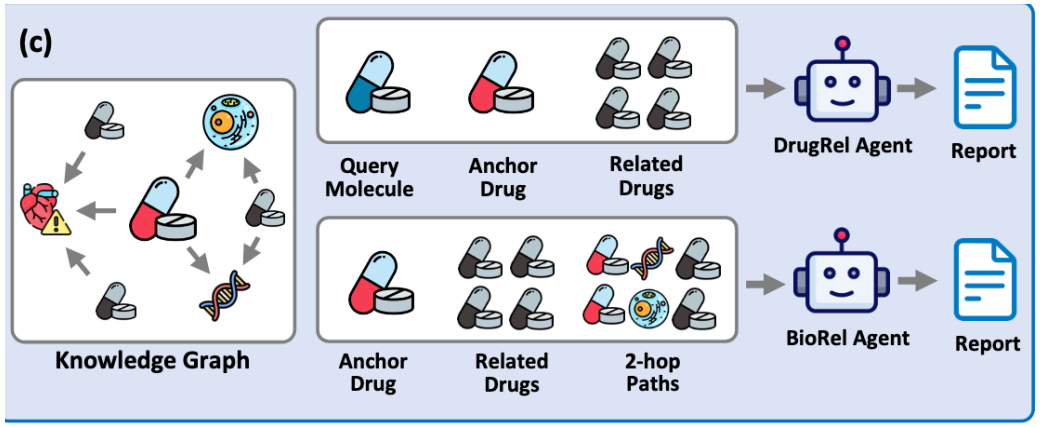
→ Multi-agent collaborative reasoning (agent가 역할 분담)



CLADD: Method





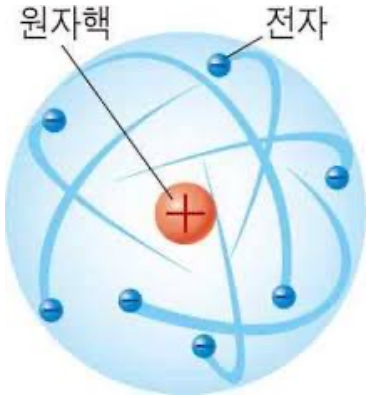


- **데이터 활용 전략 결정**: Query molecule과 task를 바탕으로 어떤 external knowledge(annotation DB, knowledge graph)를 사용할지 판단
- **MolAnn Planner**: annotation DB에서 가져온 정보가 충분한지 평가하고, 부족하면 captioning tool을 추가로 사용 여부 결정
- **KG Planner**: query molecule과 가장 구조적으로 유사한 anchor drug를 찾고, 그 유사도를 기반으로 KG정보를 사용할지 결정
- **KG 기반 맥락 정보 수집**: query molecule과 구조적으로 유사한 anchor drug를 중심으로 KG에서 관련 생물학적 정보를 탐색
- **DrugRel Agent**: anchor drug와 related drugs를 활용하여 query molecule과 유사 약물 간의 관계 및 가능성 있는 약물 작용을 분석
- **BioRel Agent**: KG의 protein, disease 등 생물학적 entity 관계를 요약하여 query molecule의 생물학적 메커니즘과 기능적 맥락을 정리

Outline

- 소재 합성 모델링
- 분광 데이터 분석
- Agentic AI for Science
- Fragmentation 기반 Molecule Representation Learning

기존 연구: Atom-level Modeling



- 분자의 물리/화학적 성질은 electron-level (전자)에서 결정됨
- 하지만, 기존 연구는 분자구조를 atom-level (원자)로 고려

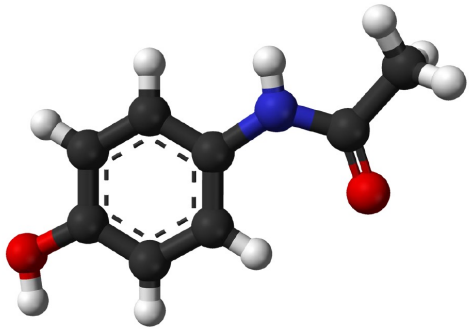
물성

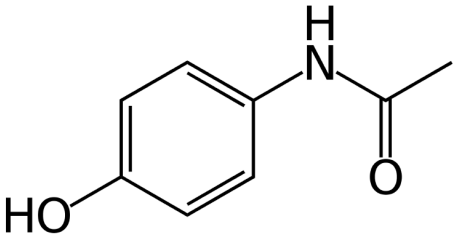
Representation learner

$$y = (f \circ g)(E)$$

분자의 물리/화학적 성질을 계산하는 함수

분자의 전자적 구조



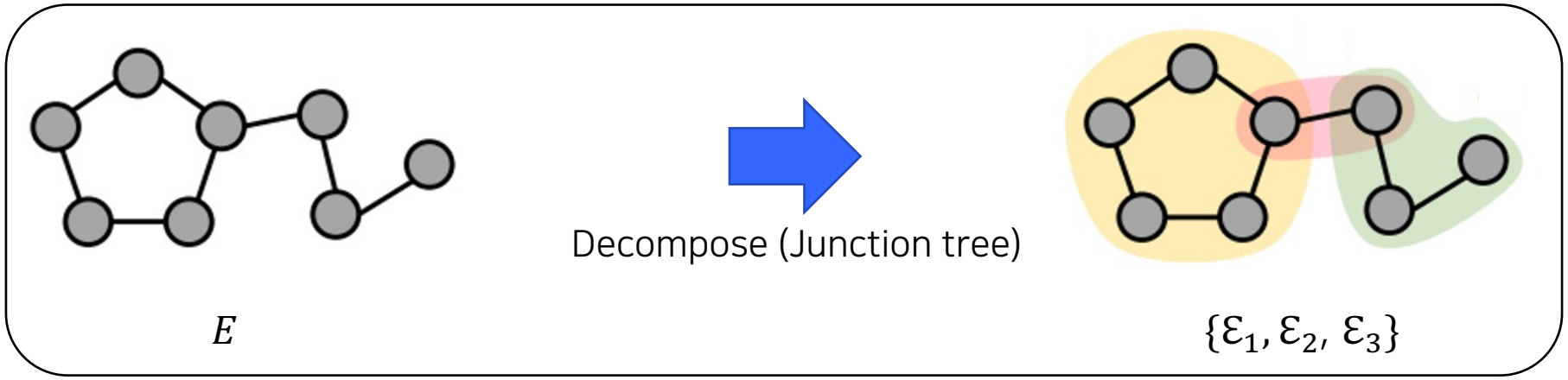


- 기존 연구는 g(E)가 분자의 atom-level 구조로 approximate이 가능하다는 가정을 하지만, 그 과정에서 정보의 손실 발생
- 전자 수준(electron-level) 정보가 분자 물성을 결정하는데, 실제 복잡한 분자에서는 전자 계산(DFT 등)이 너무 비쌈

전자수준 계산 없이 전자수준 정보를 근사하여 분자 물성 예측에 활용할 수 있을까?

Fragmentation-based Molecule Representation Learning

- 핵심 Idea: 작은 분자의 전자 정보를 활용해서 큰 분자를 이해하자



Representation learner

$$y = (f \circ g)(E)$$

분자의 물리/화학적 성질을 계산하는 함수

전자 구조

Order-invariant Representation learner

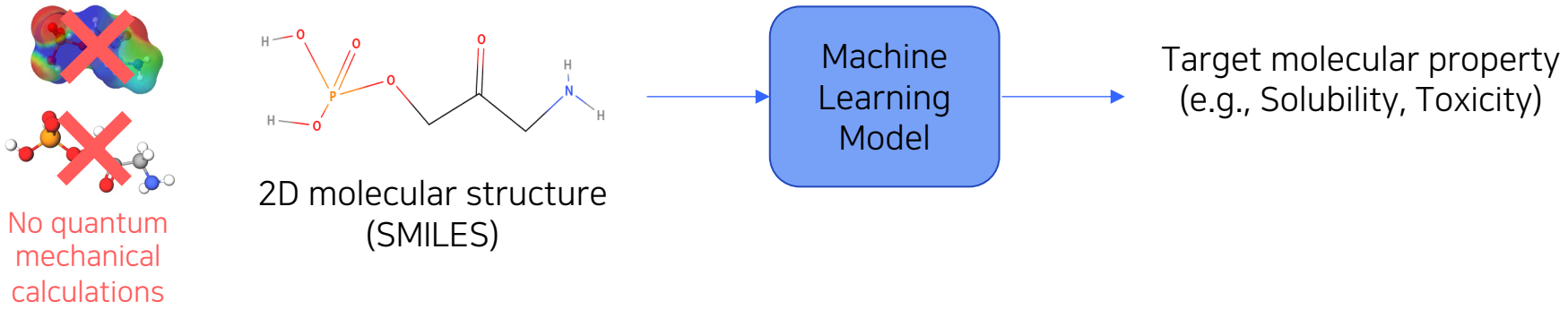
K-th electronic substructure

$$y = (f \circ g_G)(\{g_L(\mathcal{E}_1), \ldots, g_L(\mathcal{E}_K)\})$$

분자의 물리/화학적 성질을 계산하는 함수

문제 정의

- 목표: **Real-world molecular property prediction (분자 물성 예측)**
  → 전자 수준(electron-level) 정보를 고려한 **molecular representation learning**
- Input: atom-level molecular graph (2D 구조)
- Output: 실험적으로 관측된 분자 물성 (solubility, toxicity, pharmacokinetics 등)



- Fragment-level에서 계산된 전자 정보를 큰 실제 분자에 확장할 수 있는가?

→ HedMoL (KDD 2025)

- Fragment-level의 전자 정보를 이용해 전체 분자의 숨겨진 전자 표현을 추정할 수 있는가?

→ DELID (ICLR 2025)

August 3–7, 2025
KDD2025

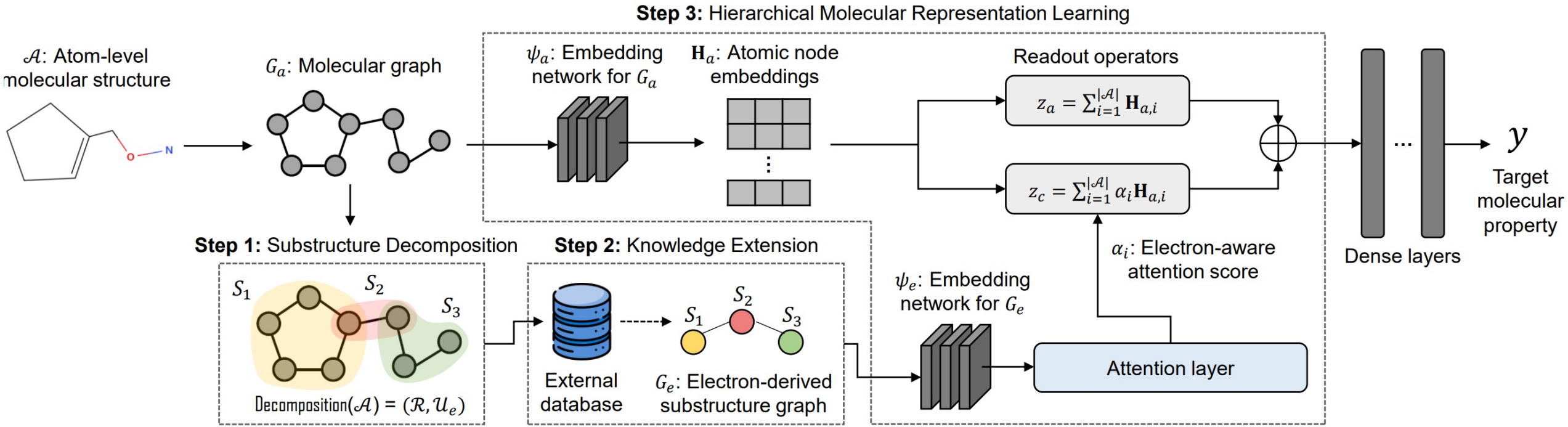
Electron-Informed Coarse-Graining Molecular Representation Learning for Real-World Molecular Physics

KDD 2025

Gyoung S. Na, Chanyoung Park

Introduction

- 핵심 아이디어: 전체 분자의 전자 특성은 부분 fragment의 전자 특성으로 근사 가능하다
  - Step 1: Substructure Decomposition - 큰 분자를 작은 fragment로 분해
  - Step 2: Electron Feature Retrieval - 각 fragment에 대해 외부 DB (QM9)에서 전자 정보를 retrieval
  - Step 3: Hierarchical Representation Learning - atom-level + electron-level representation을 결합

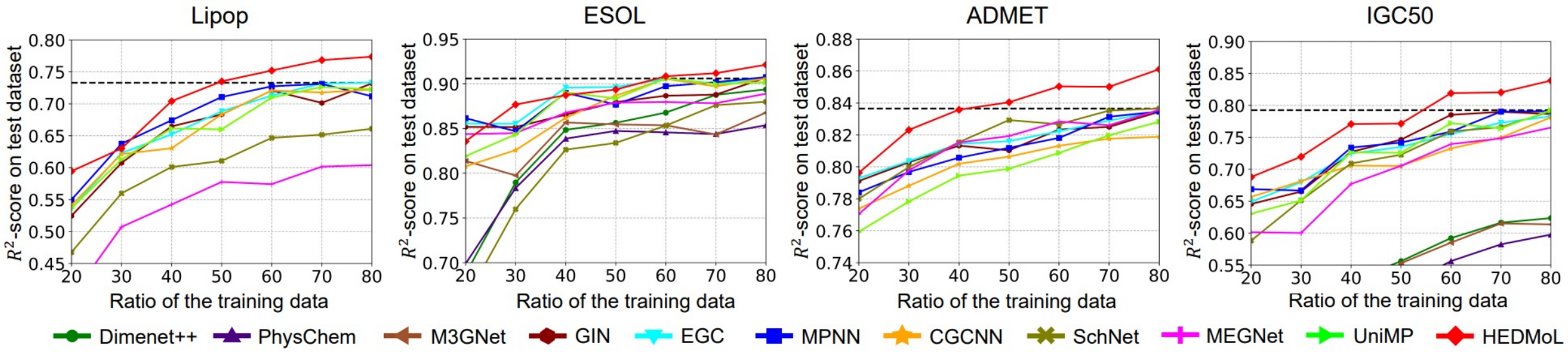


QM9 데이터의 Electron-level feature

Category	Feature Name	Unit
Energy-related feature	HOMO	eV
	LUMO	eV
	HOMO-LUMO gap	eV
	Zero point vibrational energy	eV
	Internal energy at 0 K	eV
	Internal energy at 298.15 K	eV
	Enthalpy at 298.15 K	eV
	Free energy at 298.15 K	eV
	Heat capacity at 298.15 K	eV
Polarity-related feature	Dipole moment	Debye
	Isotropic polarizability	Bohr^3
	Electronic spatial extent	Bohr^2
Other features	Rotational constant A	GHz
	Rotational constant B	GHz
	Rotational constant C	GHz

Result

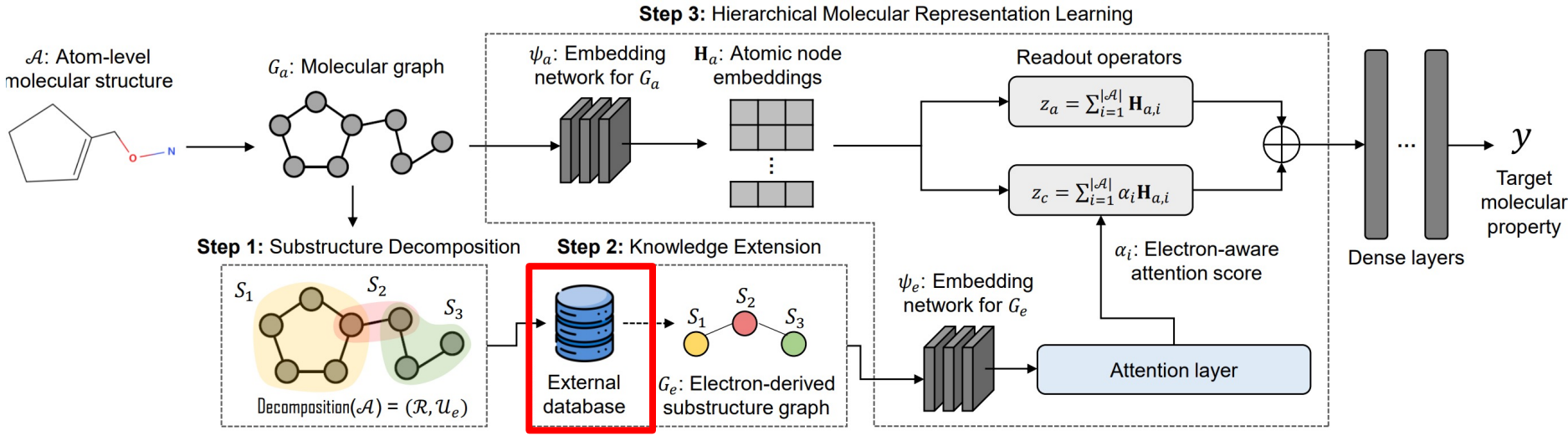
▪ 학습 데이터 크기를 달리하며 예측 정확도를 평가



• **학습 데이터가 적은 구간(20~40%)에서도 HEDMoL이 대부분의 모델보다 높은 R^2 성능을 유지함**

→ 데이터가 부족한 상황에서도 전자 정보 확장이 효과적으로 작동함을 보여줌.

Result



A subset of QM9 (molecules with six or fewer atoms)

Dataset	$C = 3$	$C = 4$	$C = 5$	$C = 6$	$C = 7$	$C = 8$
Lipop	0.732 (0.037)	0.723 (0.053)	0.735 (0.047)	0.736 (0.030)	0.738 (0.040)	0.736 (0.037)
ESOL	0.914 (0.018)	0.911 (0.016)	0.915 (0.015)	0.915 (0.016)	0.916 (0.015)	0.914 (0.019)
ADMET	0.867 (0.007)	0.862 (0.015)	0.868 (0.012)	0.861 (0.010)	0.867 (0.012)	0.865 (0.014)
IGC50	0.829 (0.017)	0.833 (0.012)	0.828 (0.016)	0.835 (0.017)	0.825 (0.018)	0.831 (0.016)

- **외부 계산 데이터베이스의 분자 크기(C=3~8)에 따라 성능 변화가 거의 없음**

→ QM9에서 작은 분자(원자 수 ≤ C)를 추출해도 예측 정확도가 전반적으로 안정적으로 유지됨.

- **대부분의 데이터셋에서 최고 성능이 특정 C에 국한되지 않음**

→ 외부 DB 규모를 크게 늘리지 않아도 충분한 전자 정보 확장이 가능함을 시사

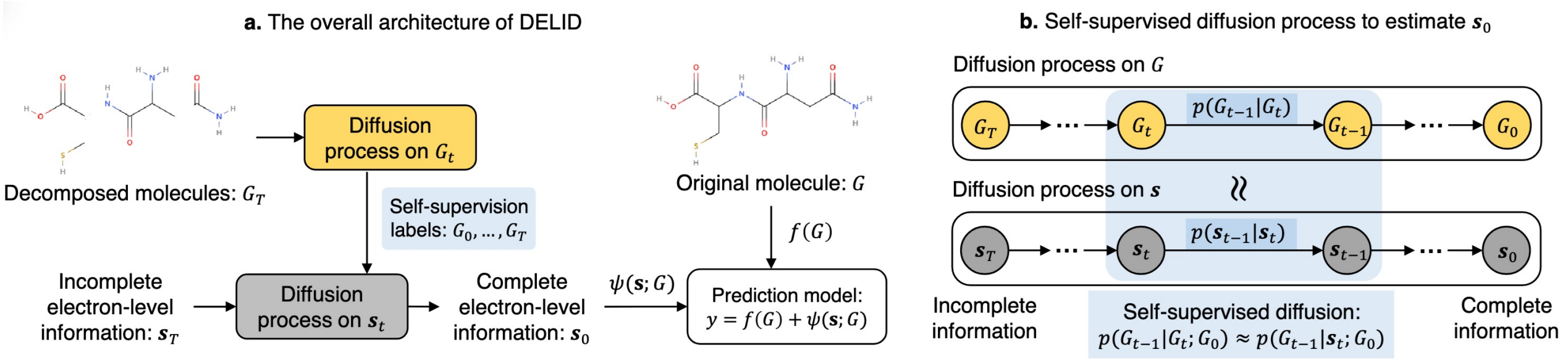
Self-Supervised Diffusion Models for Electron-Aware Molecular Representation Learning

ICLR 2025

Gyoung S. Na, Chanyoung Park

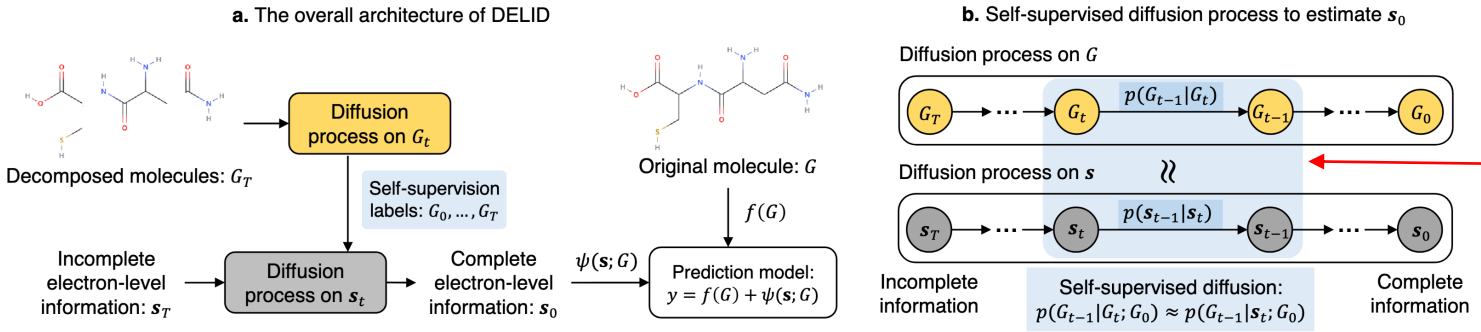
Introduction

- 관측된 fragment 전자 정보를 기반으로, 전체 분자의 잠재적 전자 상태를 diffusion으로 복원
  - Step 1: Substructure Decomposition - 큰 분자를 작은 fragment로 분해
  - Step 2: Electron Feature Retrieval - 각 fragment에 대해 외부 DB에서 전자 정보를 retrieval
  - Step 3: Self-Supervised Diffusion – Diffusion 방식으로 전체 분자의 전자정보 복원



Method

- Motivation
  작은 fragment의 전자 정보는 일부만 알려주며, 전체 분자의 전자 상태는 더 복잡함 → 단순히 fragment 정보들을 붙이면 부족함
- 불완전한 (noisy) fragment-level 전자 정보($\boldsymbol{s}_T$)를 점진적으로 정제해서, 전체 분자(G_0)의 전자 상태($\boldsymbol{s}_0$)를 복원하는 과정
- DELID는 forward-electron diffusion 없이도, graph diffusion을 teacher로 삼아 KL 정렬을 통해 reverse-electron diffusion을 학습



"전자 상태는 구조 복원 과정과 연결되어 있다"
→ 전자 상태는 구조를 설명할 수 있어야 한다

- Forward Process ($\boldsymbol{s}_0 \rightarrow \boldsymbol{s}_T$ / $\boldsymbol{G}_0 \rightarrow \boldsymbol{G}_T$)
  - 개념적으로만 정의됨
  - $\boldsymbol{s}_0$: 전체 전자 정보 (latent) / $\boldsymbol{s}_T$: Fragmented 된 전자 정보 (관측)
  - $\boldsymbol{G}_0$: 원래 분자 (관측) / $\boldsymbol{G}_T$: Fragment decomposition 결과 (관측)

- Reverse Process ($\boldsymbol{s}_T \rightarrow \boldsymbol{s}_0$)
  - 전자 상태를 직접 복원하는 게 아니라 그래프 복원을 설명할 수 있는 latent를 학습

$$\sum_{t=1}^{T} \log p(\mathbf{s}_{t-1}|\mathbf{s}_t, G) \geq \sum_{t=1}^{T} E_{q(G_{t-1}|G_t)} [\log p(\mathbf{s}_{t-1}|\mathbf{s}_t, G_{t-1})] - \sum_{t=1}^{T} D_{\mathrm{KL}}(q(G_{t-1}|G_t)||p(G_{t-1}|\mathbf{s}_t))$$

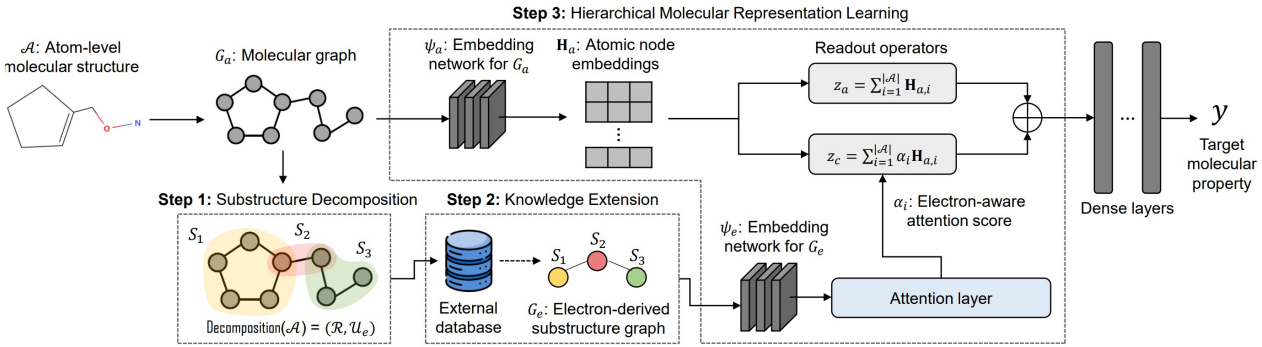
  - graph transition이 정답 분포
  - electron latent 기반 transition이 이를 모사

HEDMoL vs. DELID

- 공통 Motivation: 전자수준 직접 계산 없이, electron-aware molecular representation을 만들자
- 공통 접근방법: fragment decomposition 이후 이들의 전자정보를 외부 DB(QM9 등)에서 추출

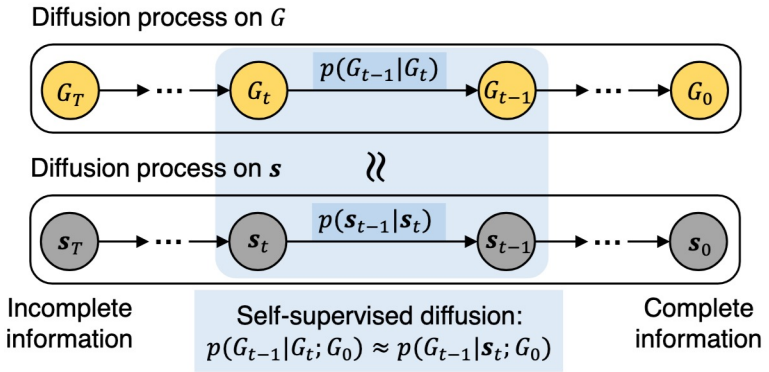
→ Fragmented 전자정보를 잘 조합해보자

HedMOL



- Fragment 기반으로 DB retrieval 이후 전자 feature를 이어 붙임 → 전자 정보는 이미 존재하는 값
- 장점: 구조가 단순하고 학습이 안정적 / 효율적
- 단점: Fragment 전자 정보에 noise 있으면 그대로 반영됨

DELID



- fragment에서 얻은 local 전자 정보를 기반으로, 전체 분자의 전자 상태를 점진적으로 정제해 가는 과정 → 전자 정보는 latent variable
- 장점: Diffusion 단계에서 global 구조와 모순되는 부분을 점진적으로 보정
- 단점: 모델 복잡도가 높음 / 학습 불안정 / 계산 비용 ↑

Summary

- 소재 합성 모델링
- 분광 데이터 분석
- Agentic AI for Science
- Fragmentation 기반 Molecule Representation Learning

경청해 주셔서 감사합니다.

Reference

- (NeurIPS 2024) Retrieval-Retro: Retrieval-based Inorganic Retrosynthesis with Expert Knowledge
- (preprint 2026) MSP-LLM: A Unified Large Language Model Framework for Complete Material Synthesis Fragmentation 기반 Molecule Representation Learning
- (ICLR 2026) IR-Agent: Expert-Inspired LLM Agents for Structure Elucidation from Infrared Spectra
- (preprint 2026) Periodic Complex Stochastic Processes for Retrieving Atomic Structures of Unknown Matters
- (preprint 2026) Machine Collective Intelligence for Scientific Discovery
- (AAAI 2026) RAG-Enhanced Collaborative LLM Agents for Drug Discovery
- (ICLR 2025) Self-Supervised Diffusion Models for Electron-Aware Molecular Representation Learning
- (KDD 2025) Electron-Informed Coarse-Graining Molecular Representation Learning for Real-World Molecular Physics