Grounding Large Language Models in Reaction Knowledge Graphs for Synthesis Retrieval

Graph&Data seminar

Lorenzo Di Fruscia / 05-02-2026

TU Delft

Chemical synthesis in our society

- Chemical synthesis is the process of planning and executing reactions to produce target molecules (medicines, materials...)[1]
- Chemical databases keep expanding every year
- Reaction planning is structured retrieval under constraints (reactants/products, conditions…)

→ Computer-assisted planning tools help us navigate this vast chemical knowledge

[1] E. J. Corey, "*Computer-Assisted Design of Complex Organic Syntheses*", Science, 1969

The advent of Large Language Models

- The last few years have seen the rise of Large Language Models[1] (LLMs)

- Large* transformers trained on large* datasets → Emergent Capabilities
  1. Knowledge retrieval
  2. Learning from examples
  3. Reasoning
  4. Tool usage
  5. …

- **In-Context Learning**: model performs tasks using natural language instructions/examples provided as input
- *Why interesting?* No parameter updates → versatile and cheaper

*Billions/trillions of tokens/parameters

[1] Zhao et al. "A Survey of Large Language Models." arXiv, (2023)

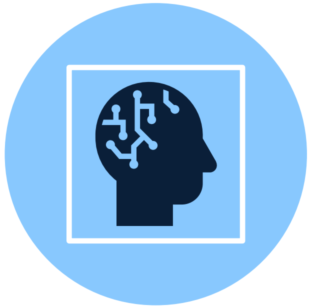
Chemical synthesis in the era of LLMs

- LLMs could aid in identifying chemical pathways from language prompts

- *However*, shortcomings of *out-of-the-box* LLM prompting include:
  - Hallucinations
  - Limited domain knowledge
  - Weak multi-step reasoning

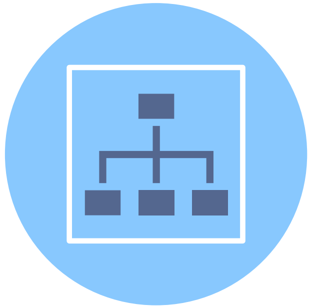
KG-Enhanced LLMs for trustworthy outputs

- A Knowledge Graph (KG) represents information as nodes (entities) and edges (relationship)
- By connecting it to a LLM we can:

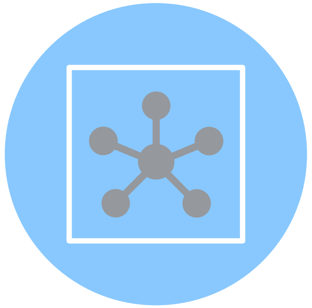


Provide structured knowledge during inference

Capture global relational & structural context



Preserve hierarchical dependencies

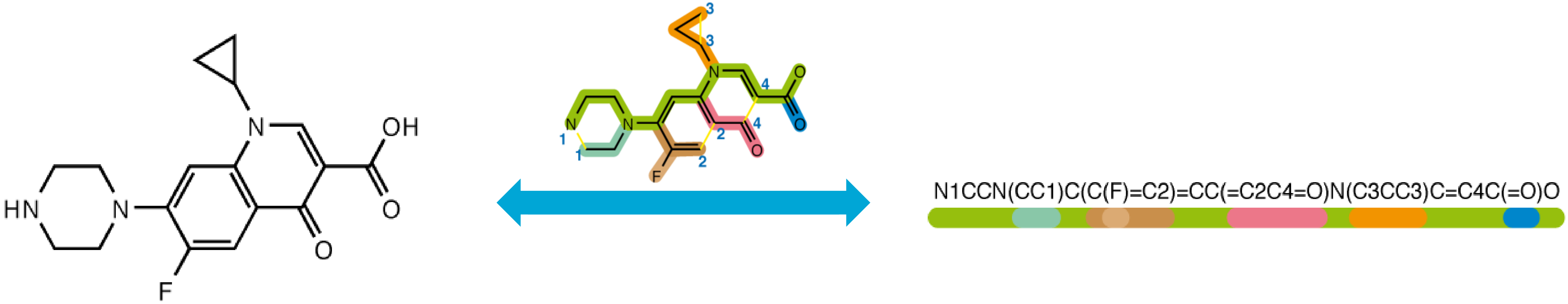


Enable multi-hop reasoning via graph traversal

Can we ground LLMs in chemical reactions databases to make their outputs trustworhty?

Machine-friendly chemistry representations

Molecules can be equivalently represented as **graphs** or **strings** (SMILES[1])



The same holds for chemical reactions!

[1] Weininger, David. “SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules.” Journal of Chemical Information and Computer Sciences 28, no. 1 (1988).

Groups of reactions can form pathways

Molecules can be involved in multiple processes, leading to multi-step reactions

Reactants **Product**

CO . CC(=O)O >> CC(=O)OC

r1 : A + B ⟶ C

r2 : C ⟶ D

r3 : D + E ⟶ F

⋮

Reaction Equations

TU Delft

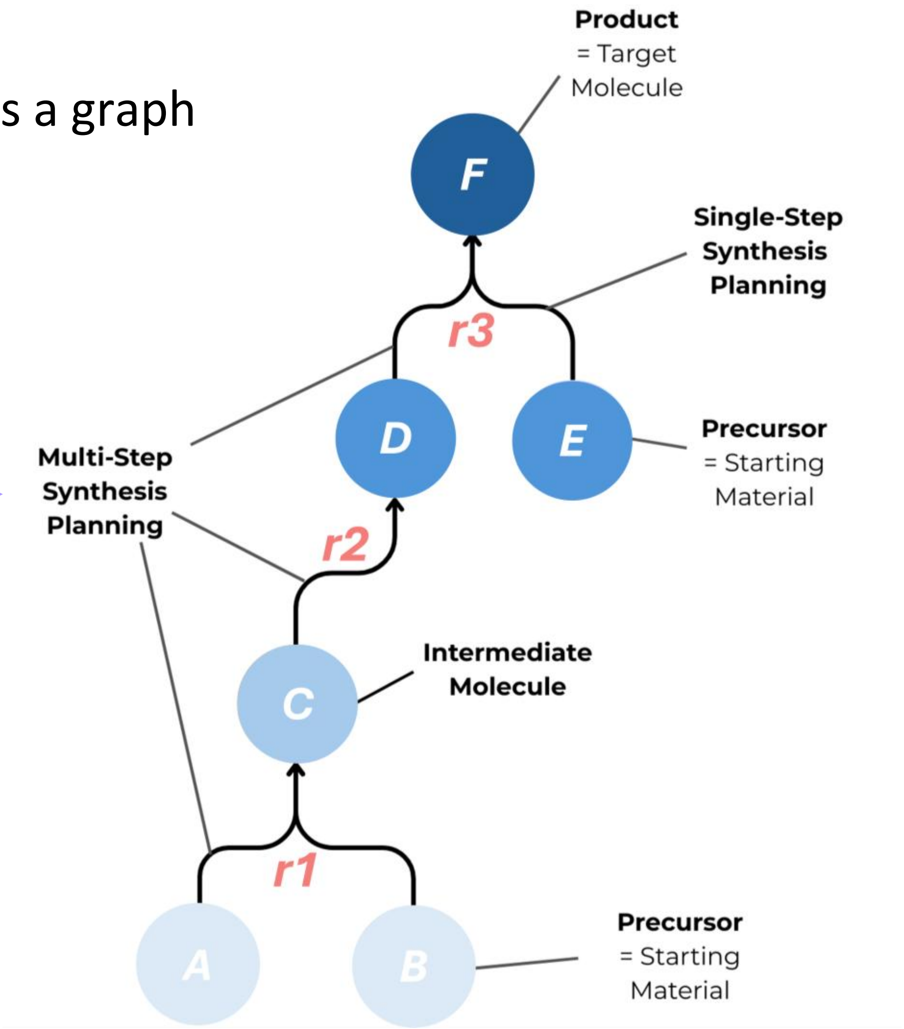
Pathways as reaction networks

- We can visualize multiple reactions as a graph

Single step

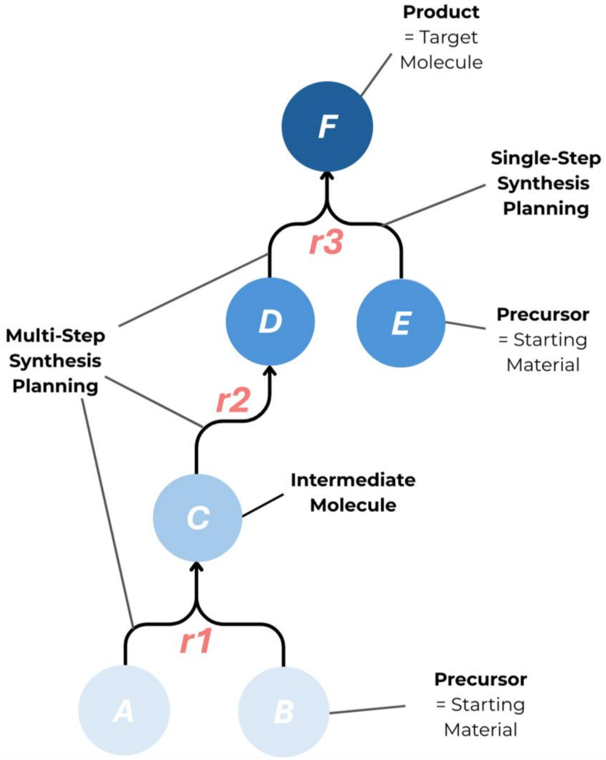
$r1$: A + B → C
$r2$: C → D
$r3$: D + E → F

Reaction Equations



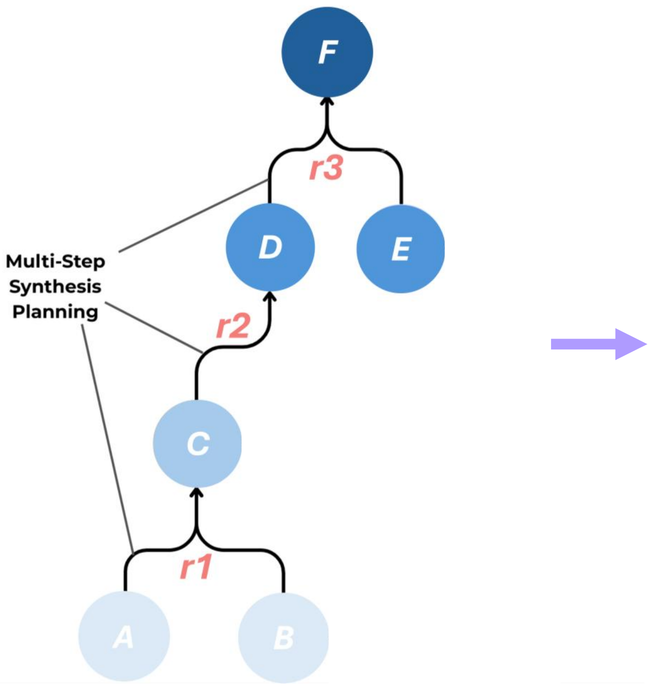
A way to preserve network connectivity

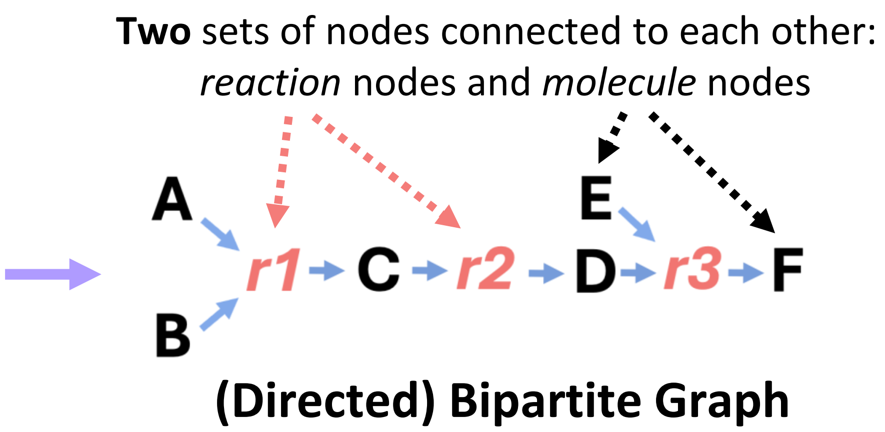
- How to efficiently store the reactions, while preserving the network topology? *Bipartite graphs*!
- Handy representation for elegant multi-component interactions



A way to preserve network connectivity

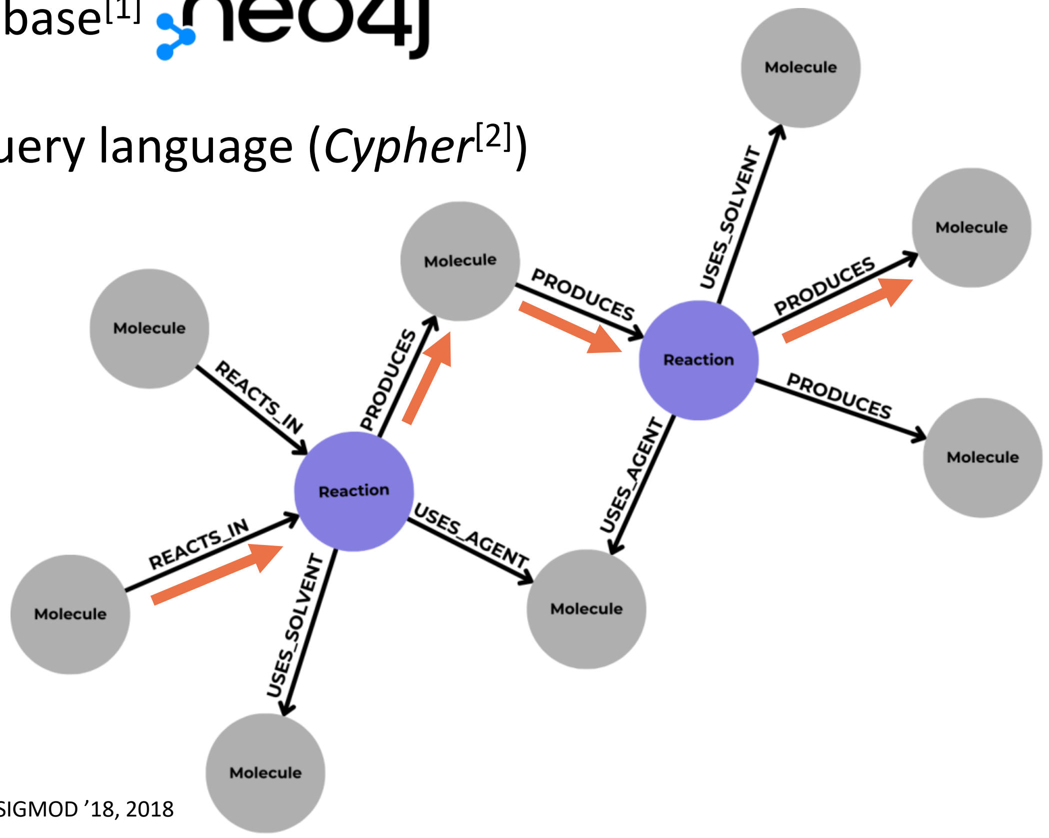
- How to efficiently store the reactions, while preserving the network topology? *Bipartite graphs*!
- Handy representation for elegant multi-component interactions





A flexible storage solution: graph database

- Chemical reactions are far from simple: *co-reactants*, *multiple-products*…
- We encode the bipartite graph with a graph database[1] neo4j
- Straightforward graph traversal through graph query language (*Cypher*[2])



TU Delft

[1] Neo4J, https://neo4j.com/
[2] N. Francis *et al.*, "*Cypher: An Evolving Query Language for Property Graphs*", SIGMOD '18, 2018

Task selection

Single-step Reactions

1. Precursor Identification
2. Product-to-Reaction Mapping
3. Condition-Based: Agents & Solvents
4. Best-Yielding Reaction
5. Multi-Product Synthesis
6. Co-product Identification

1200 data samples in total
200 per task

Multi-step Reactions

1. Multi-Step Precursor Discovery
2. Pathways to Target Molecule
3. Pathways Between Molecules
4. Reaction Pathway Intermediates

1200 data samples in total
300 per task

Dataset: chemical reactions are sampled from a 50k reactions subset of the USPTO database[1]

[1] Lowe, D. Chemical reactions from US patents (1976-Sep2016)

Prompt structure for the LLM

- Each sample is framed as a natural language-for-retrieval task
- Each sample is associated to multiple prompts of increasing complexity

LLM: GPT-4.1[1]

You are an expert in generating Cypher queries for a Neo4j knowledge graph with the following schema: [...] based solely on user input.

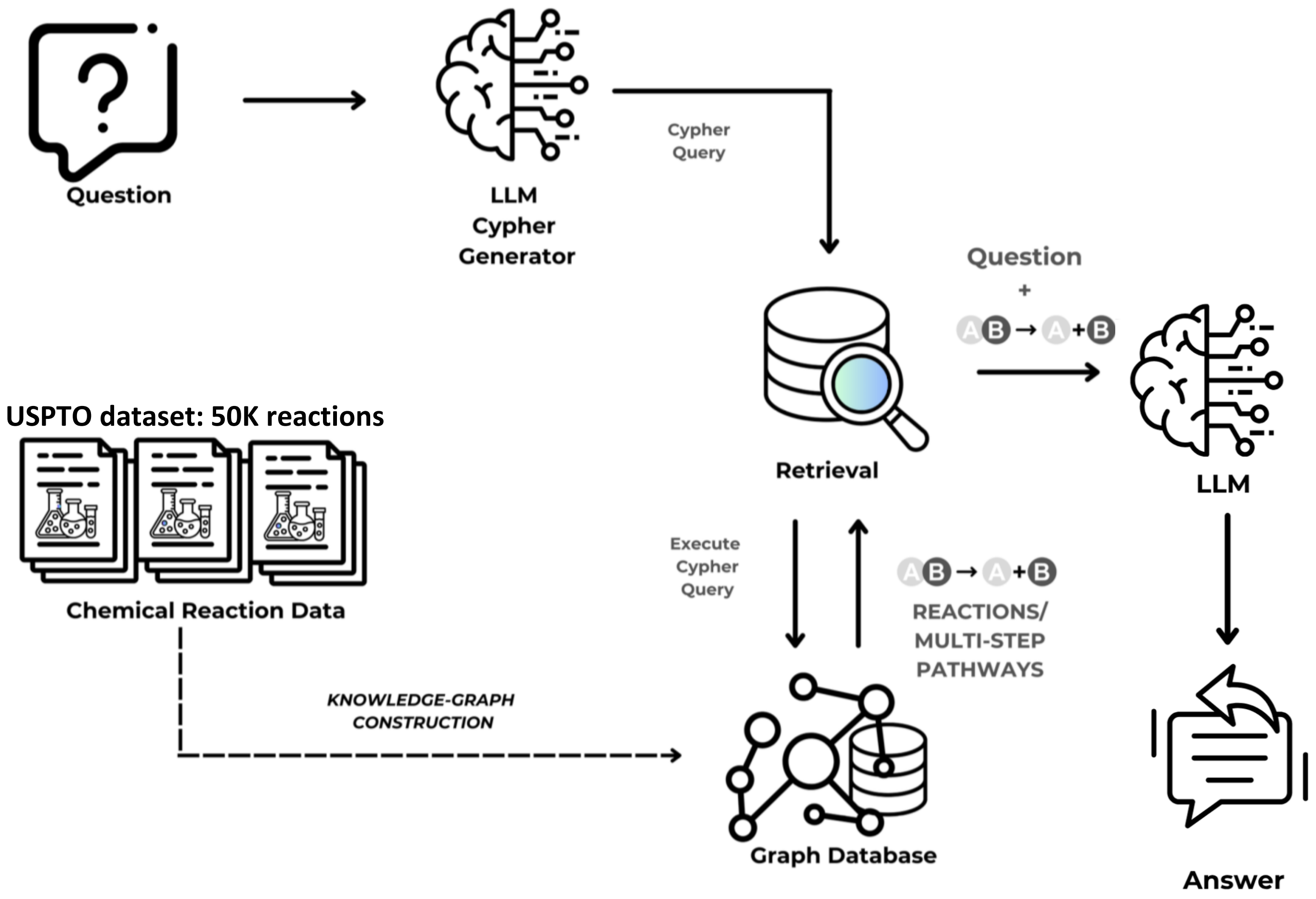
1. General guidelines: [...]
2. Always use the following methods to bind the variables [...]
3. Ensure that the directionality of every reaction is correct [...]
4. Additionally report the following contextual entries [...]
5. Follow the overall scheme for a correct retrieval pattern [...]

prompt complexity

TU Delft

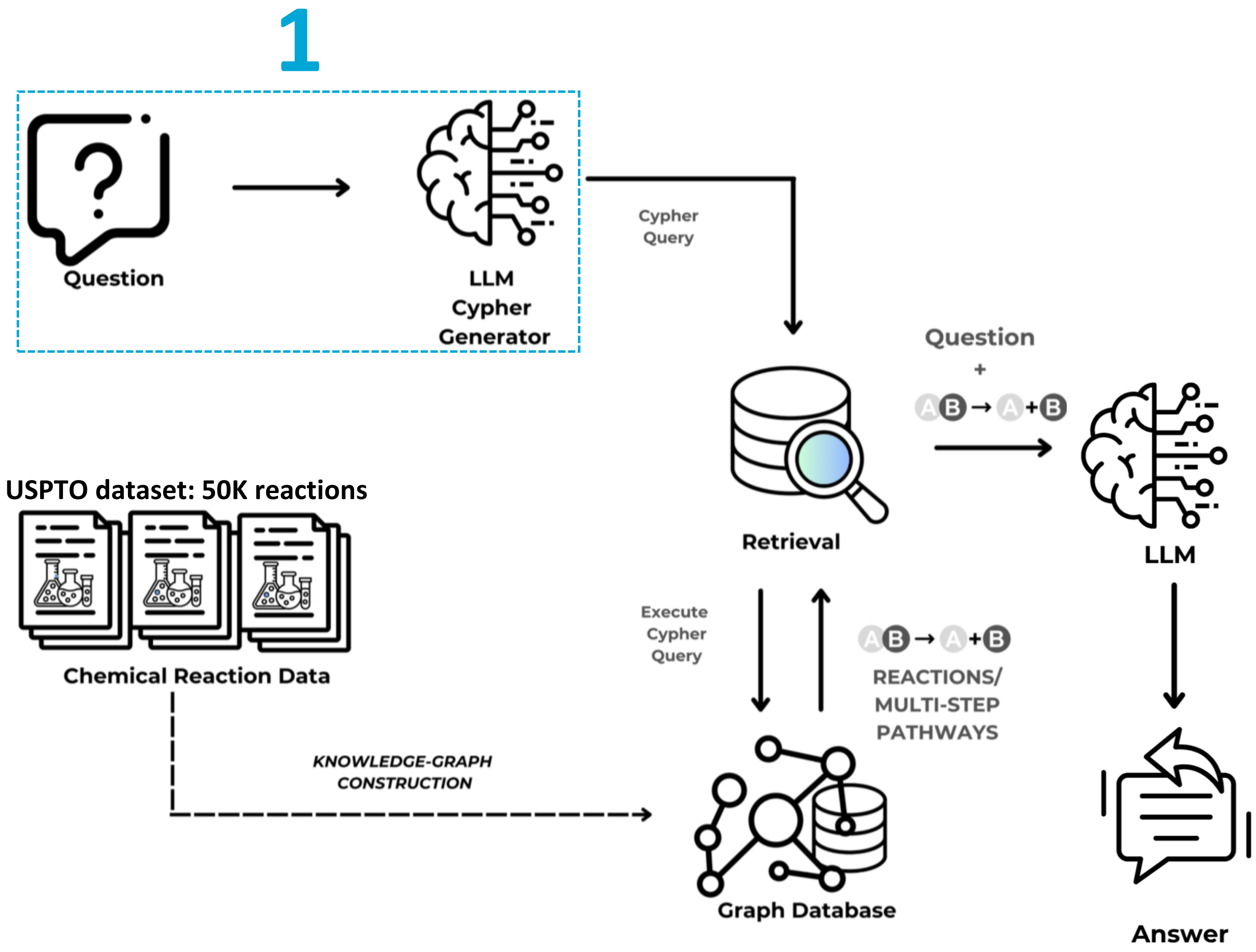
[1] GPT-4.1-mini-2025-04-14

The broad picture



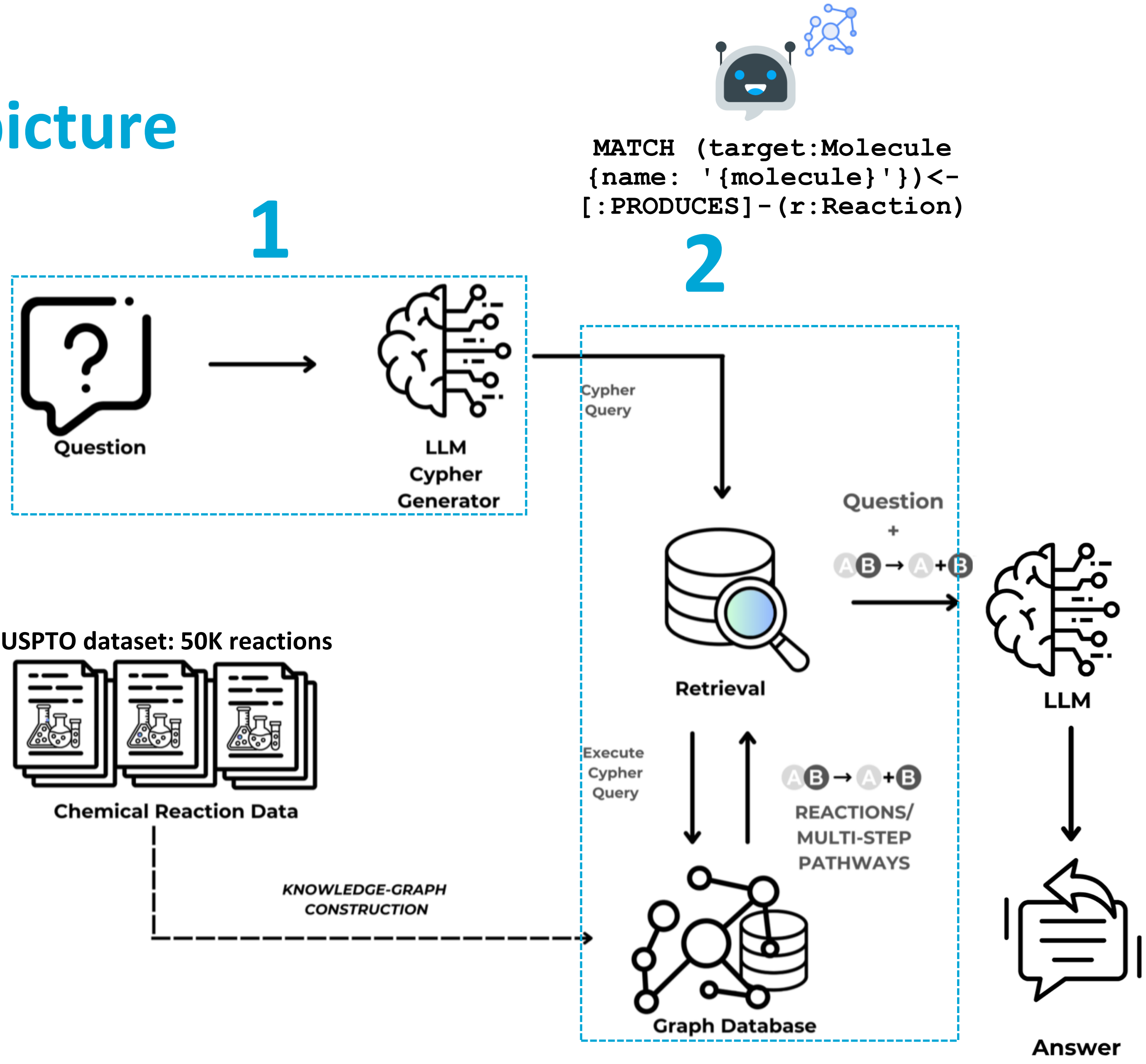
The broad picture

What are the direct precursors used to synthesize {molecule}?

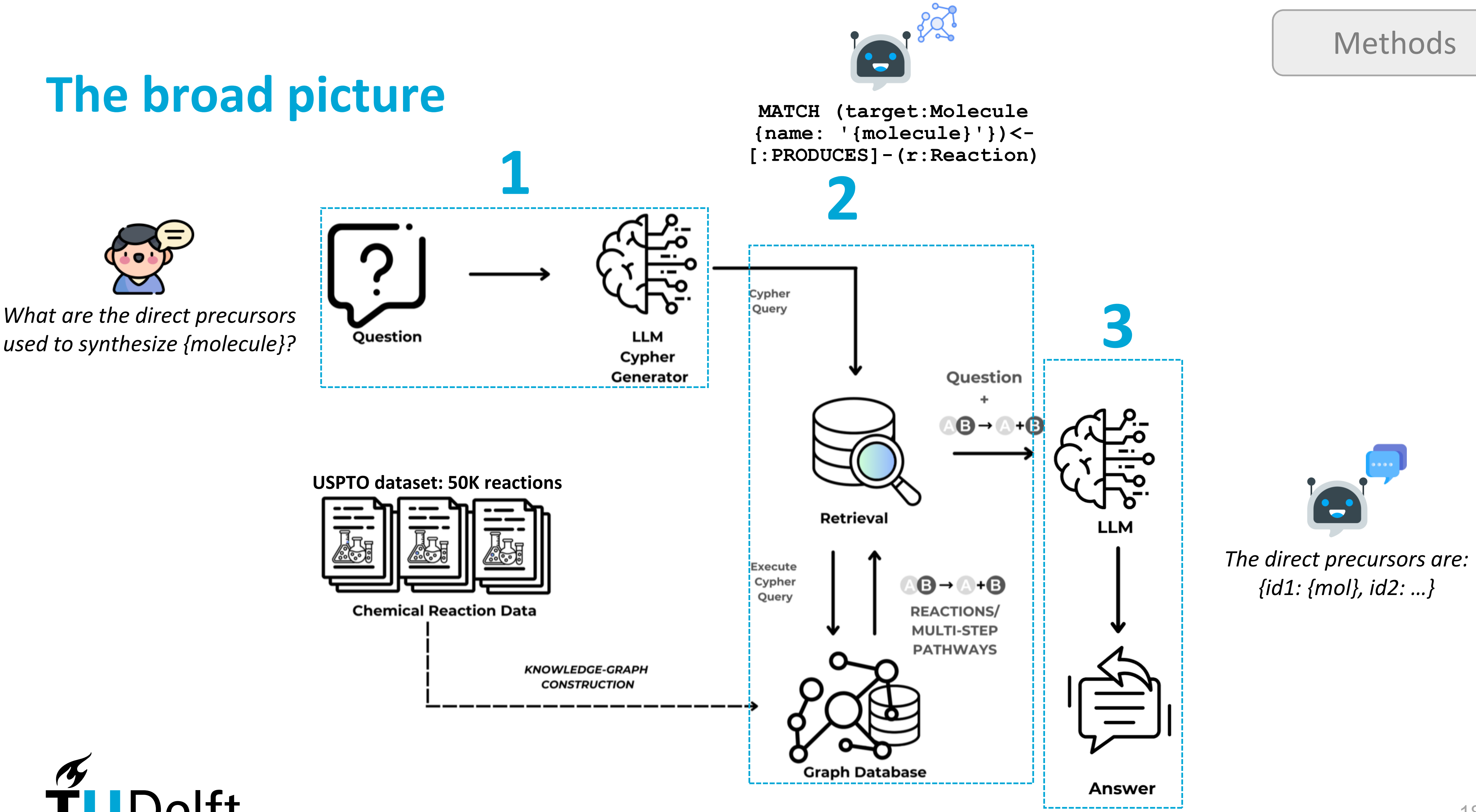


The broad picture

What are the direct precursors used to synthesize {molecule}?



The broad picture

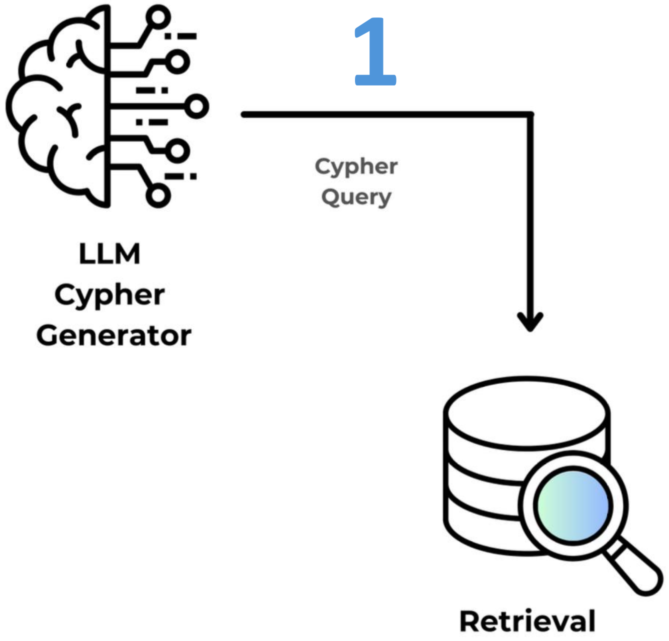


TU Delft

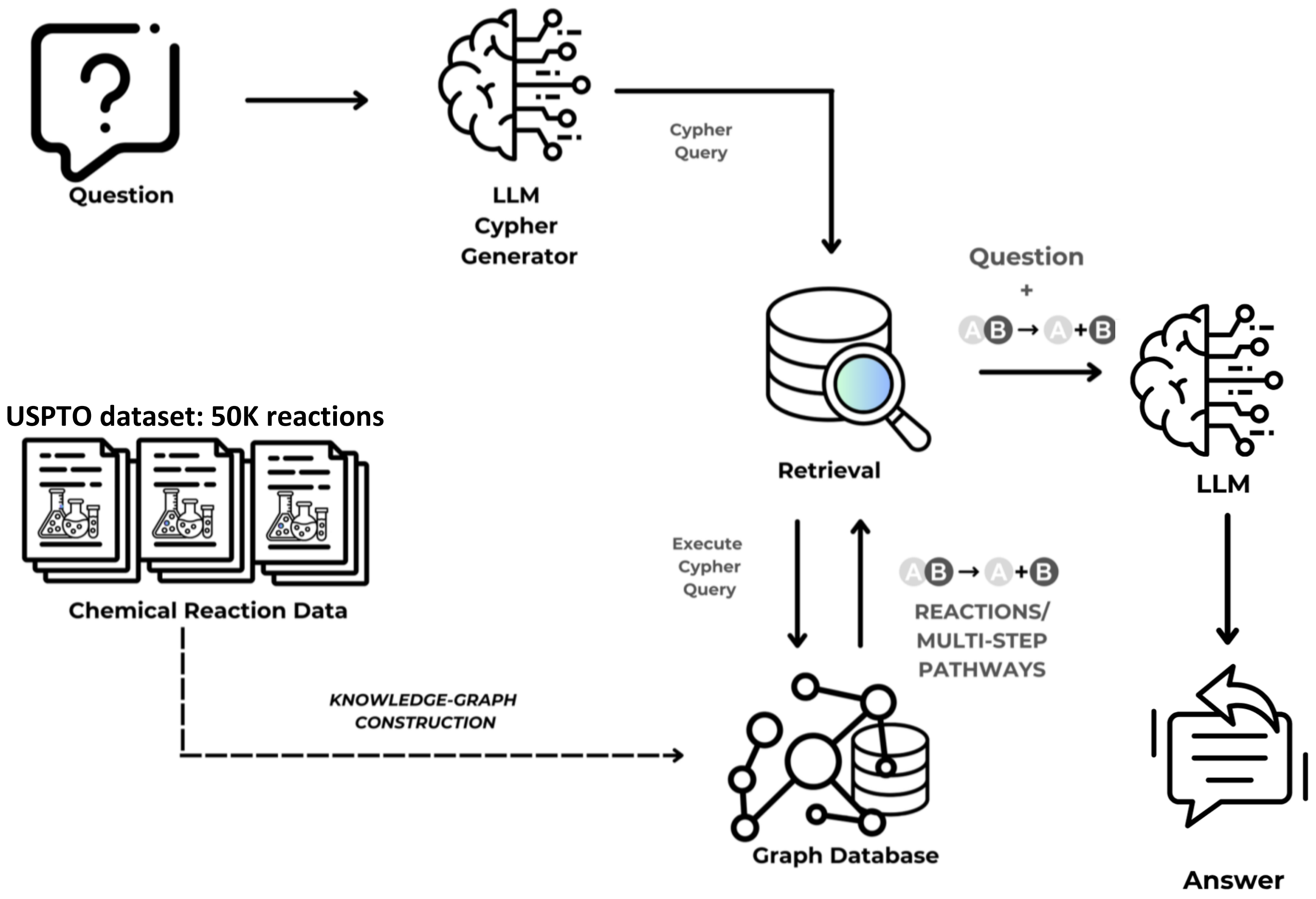
Two step evaluation: Text-to-Text (query) generation

- *How similar is the LLM-generated query to the ground truth query?*
- Metrics: **Bleu, Meteor, Rouge-L** ∈ [0,1]

Word overlap,
longest common sequence,
synonym matching,
…



The broad picture



TU Delft

Two step evaluation: Retrieval from executed queries

- *Can the LLM interpret the query output and provide a structured answer?*
- Metrics: **Keys** (e.g., reaction_id) with Precision, Recall, F1, and **path** (partial or exact) matching

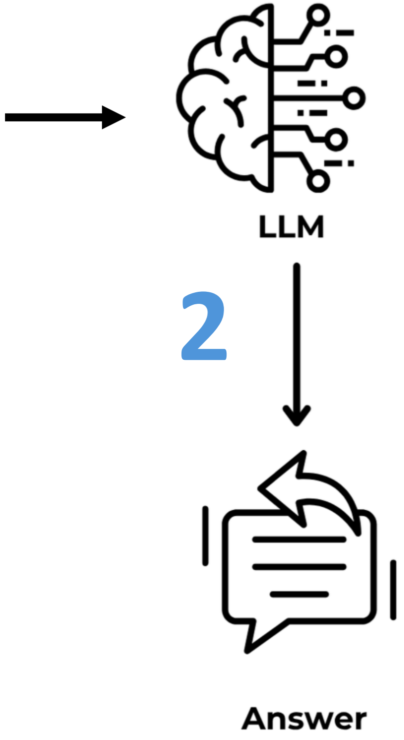
Single-Step retrieval format

```
{
    "reaction_id": id,
    "products":  [SMILES1],
    "reactants": [SMILES2, SMILES3],
    "agents":    [SMILES4, SMILES5],
    …
}
```

Multi-Step retrieval format

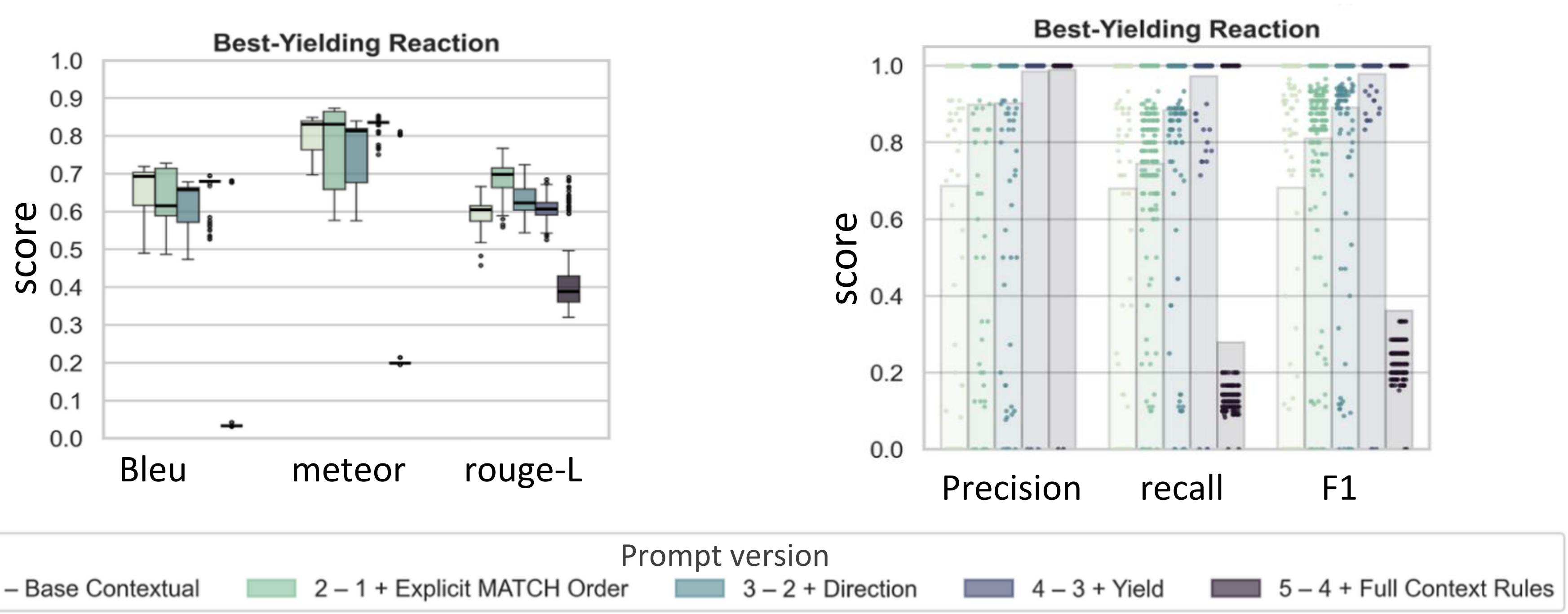
```
{
    "reaction_nodes": [
    {"id": id1, "type": "Reaction"}
    {"id": id2, "type": "Reaction"}
     …]
}
```

Query output



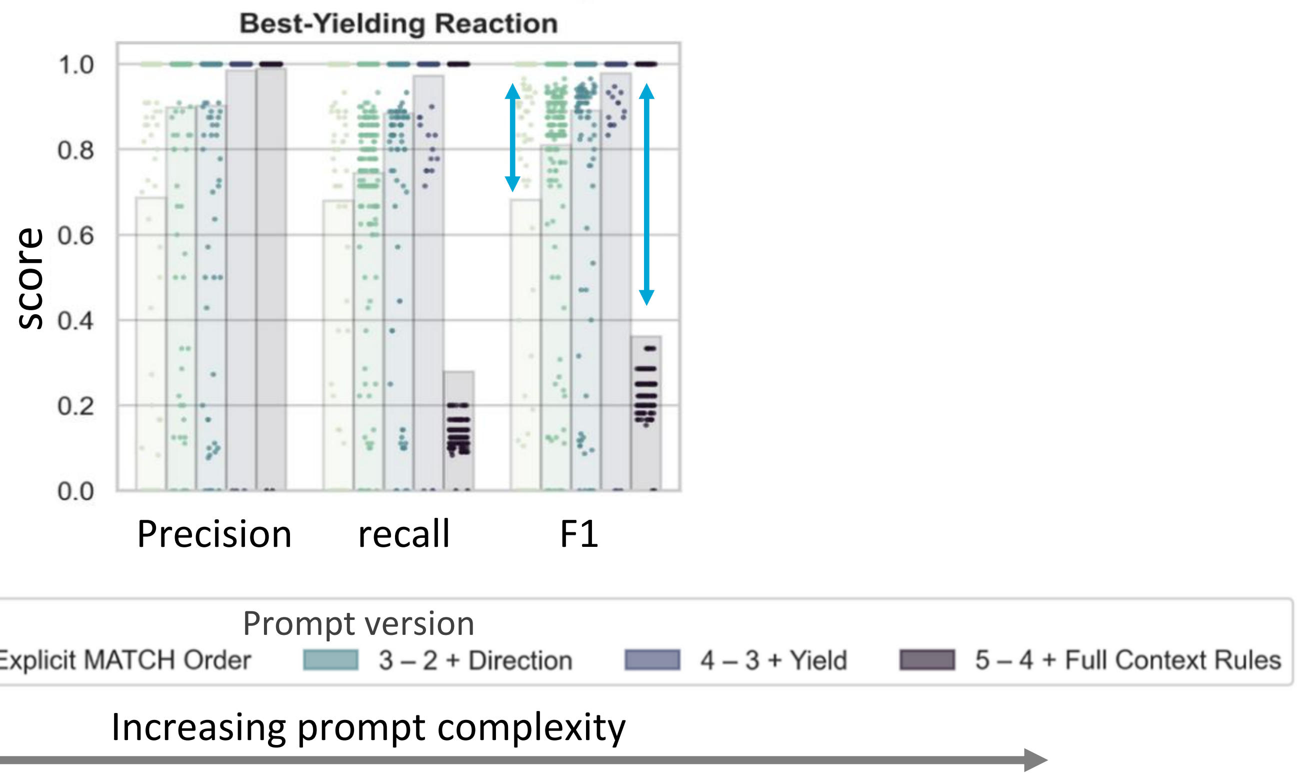
High query similarity =/= correct retrieval

- Cypher queries can be semantically similar to reference queries, while altering the retrieval
- Retrieval-level metrics reveal the real errors: wrong products, missing reactants, wrong paths…



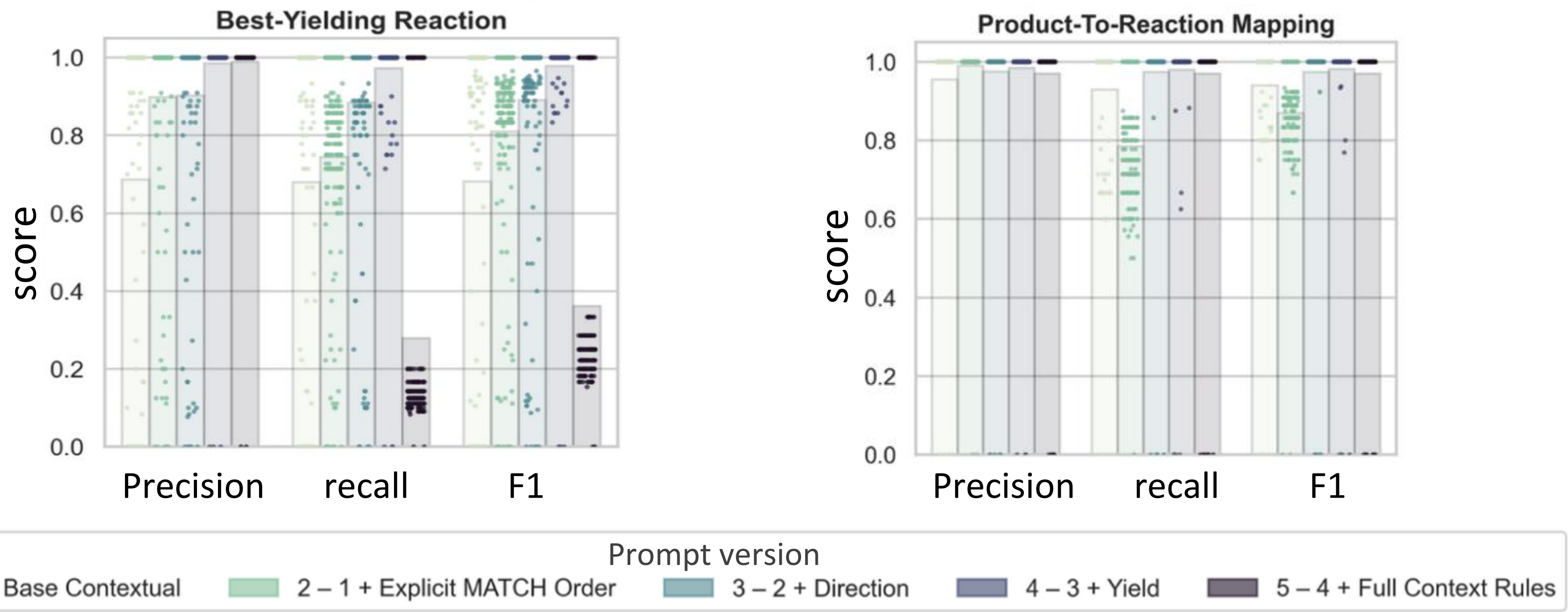
Prompt complexity is important

- Prompts that are too unconstrained or over-constrained can harm retrieval



Prompt complexity is important

- Prompts that are too unconstrained or over-constrained can harm retrieval
- *However,* prompt strategy is also task sensitive

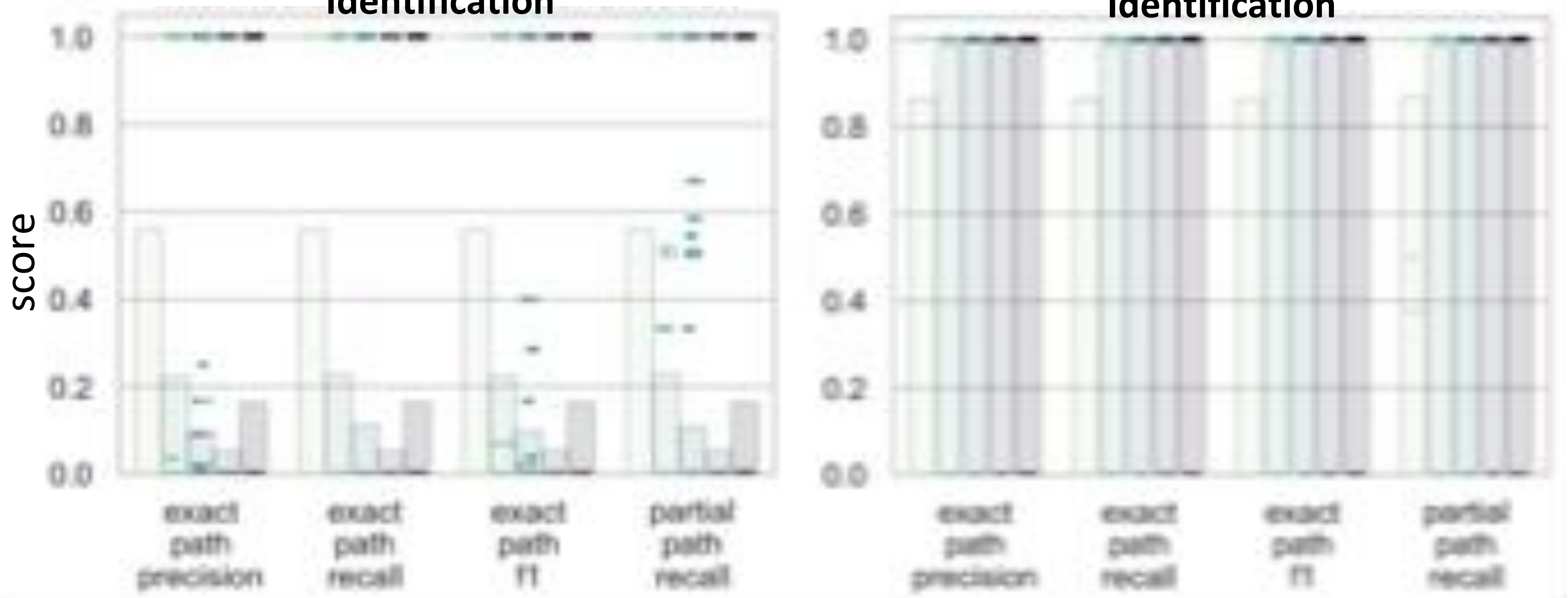


Increasing prompt complexity →

TU Delft

Examples are important for multi-step graph traversal

Intermediate Molecule Identification | **Intermediate Molecule Identification**



Zero-shot | One-shot

TU Delft

One shot prompting and retrieval: a closer look

- Even a single example greatly reduces retrieval errors, mostly due to incorrect graph traversal in the query

Zero-shot prompts

One-shot prompts

Conclusions

Our results provide practical guidelines for reaction planning and database retrieval with LLMs

- We formulate reaction/pathway retrieval as Text2Cypher over a reaction knowledge graph
- *Query similarity can be misleading:* evaluation should be done at the retrieval level (products/reactants/paths)
- *Prompting is task sensitive:* more complexity is not always better
- *One example is a big lever*: one-shot prompting greatly reduces multi-step graph traversal mistakes

Acknowledgments

- Olga Bunkova
- Sophia Rupprecht
- Artur M. Schweidtmann
- Marcel J.T. Reinders
- Jana M. Weber

The Delft Bioinformatics Lab

Accepted at ML4Molecules 2025
(ELLIS UnConference workshop)

TU Delft

Thank you for your attention!