MULTIMODAL RAG FOR ACCELERATING NOVEL CATALYST DISCOVERY

Presented by Aishwary Sharma: 2022CH11437
Supervisor: Prof. Manoj Kumar Ramteke
Dept. of Chemical Engineering, IIT Delhi

Overview

- The Problem
- Novelty
- Objective
- Methodology
- Results
- Conclusion
- Future work

The Problem

Scientific literature is multimodal, but our tools are not

- Catalyst research and design now increasingly depends on DFT (Density functional theory) analysis and advnacment s in ML based methods

Critical descriptors are hidden inside:

- **Tables**
- **Plots (DFT curves, volcano plots)**
- **Heatmaps & performance charts**

These can be called **dark data modalitites** as:

- Modern LLMs only extract text
- The image content is ignored unless described in text
- The tables are either misread or not extracted
- There is no scope for domain fine-tuning

Novelty: Multimodal RAG

RAG solves hallucination; multimodal RAG solves extraction.

- Retrieval-Augmented Generation (RAG) grounds LLM answers in retrieved context
- Multimodal RAG (MRAG) extends retrieval to:
  - Images
  - Figures
  - Tables
  - Diagrams
- **Biomedical field** already uses **MRAG** (BioMol-MQA, AlzheimerRAG)
- But no multimodal RAG system exists for catalysis

Our Objective

Parse scientific PDFs into text, tables, figures

Embed everything into a unified vector space

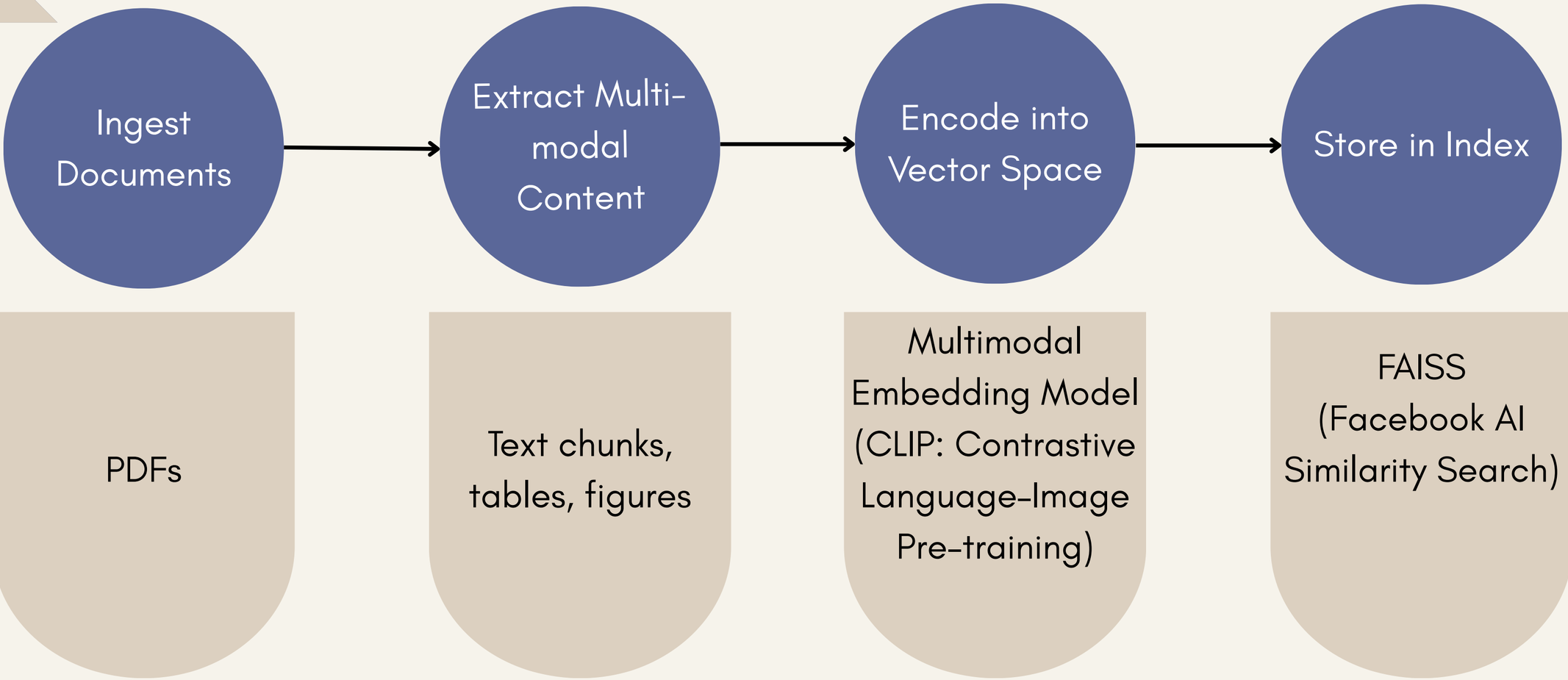
Retrieve relevant evidence across modalities

Generate grounded, traceable answers through a conversational interface

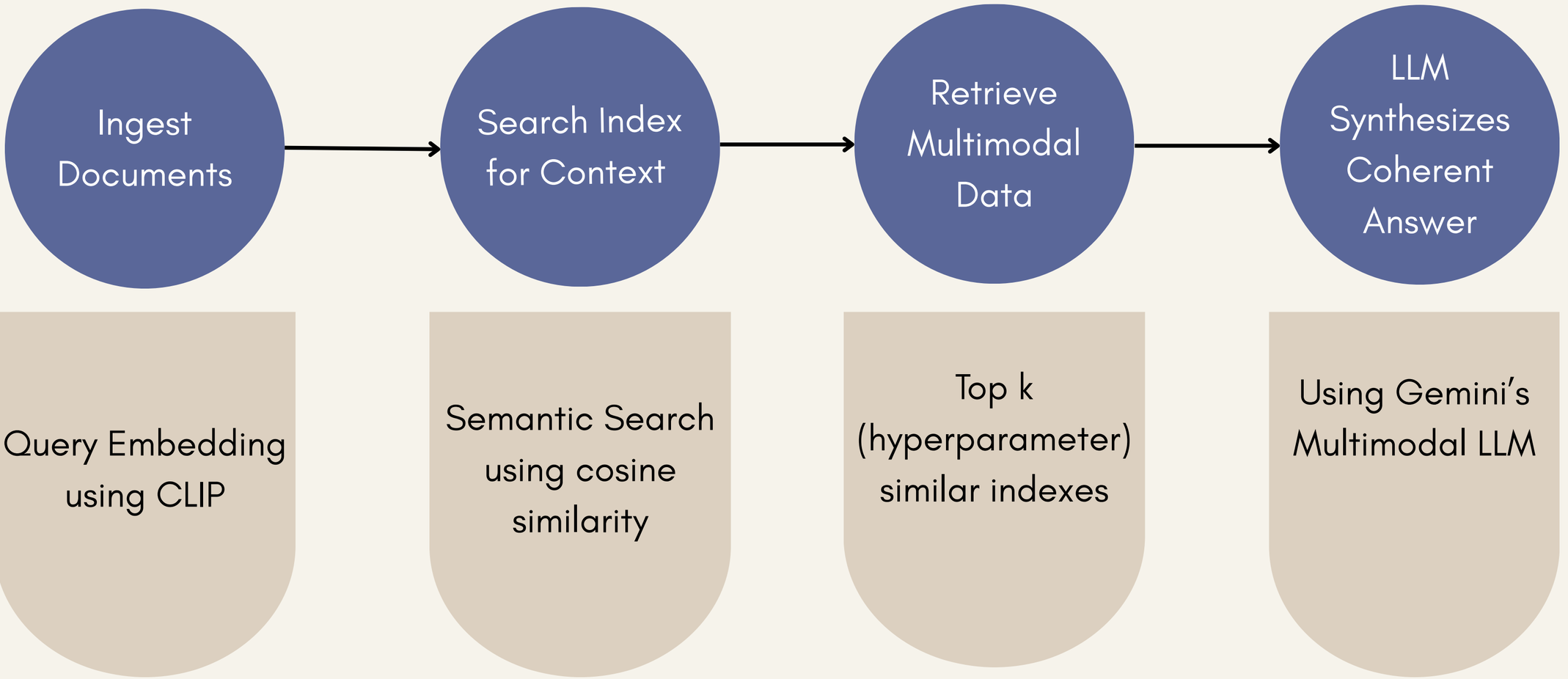
METHODOLOGY

System architecture: Two phases

Phase-1: Indexing



Phase 2: Retrieval & Generation



Extracting & Retrieving Images from Scientific PDFs

The Challenge:

- CLIP image embeddings are different from text embeddings
- During cosine-similarity retrieval, image vectors rarely match text queries

The Solution:

Metadata-Driven Image Retrieval

- For each image, we extracted rich textual metadata:
  - ***Figure caption detection*** (page segmentation + nearest-box heuristic)
  - ***OCR text*** inside the figure (axes labels, legends, units)
- Used only the meta-data embedding for the retrieval of images
- Used the corresponding actual images in the prompt while generation

Why Metadata Works Better

- Text and images occupy very different regions of CLIP's embedding space
- Metadata brings images into the same semantic space as the query
- Achieved significant improvement in Accuracy@8 **(from 0.3 to 0.7)**

Results

Retrieval Performance across Queries

Evaluated on 40 queries (20 text-based, 20 image-based)

Metric: Accuracy@8 which tells whether the correct supporting chunk appears in top-8 retrieved neighbors

Query type	No. of queries	Accuracy @8
Text-based	20	0.85
Image-based	20	0.7
Overall	40	0.78

Table: Retrieval performance across text and image queries

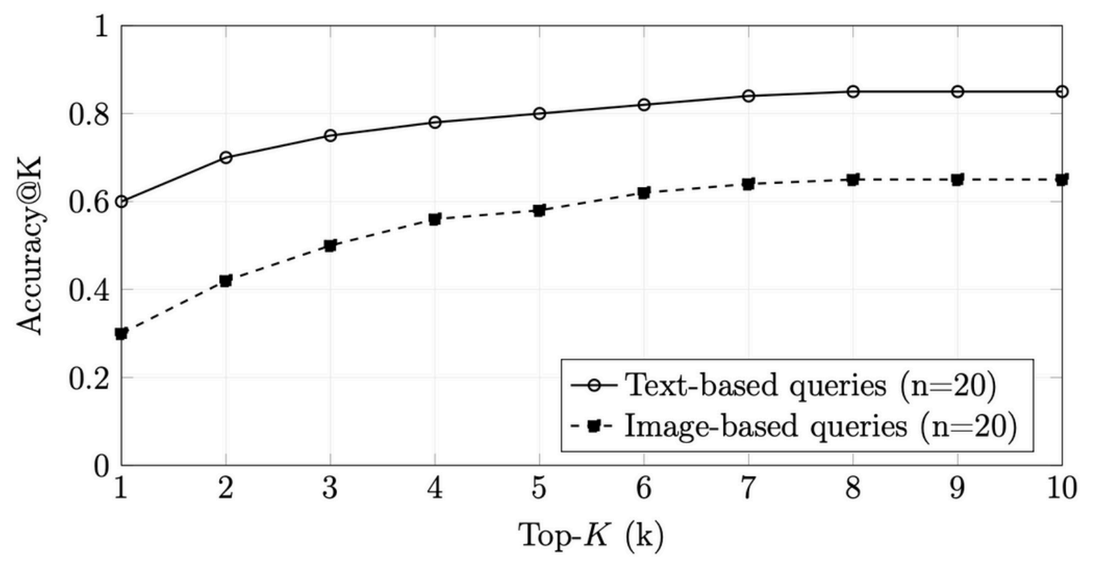


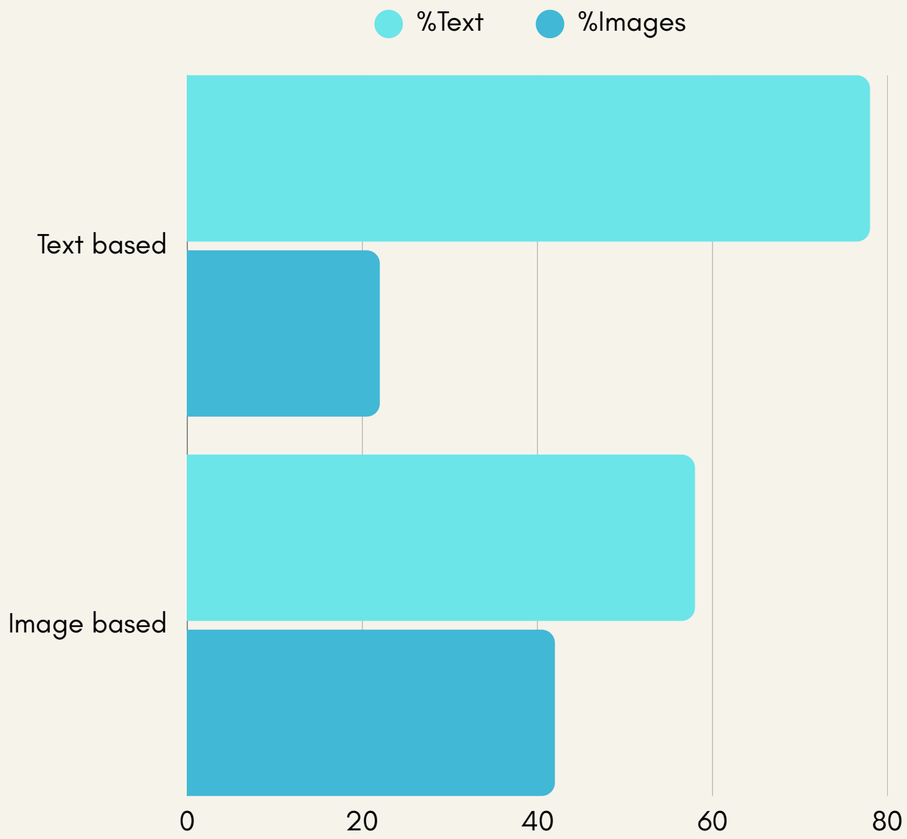
Fig: Accuracy@k for text-based and image-based queries

Modality Sensitivity Analysis

This experiment checks whether the retrieval system retrieves the correct modality of data

Modality Breakdown of Top-8 Retrieved Chunks

- Image based: 42% retrieved chunks were images
- Text based: 78% retrieved chunks were text
- Confirms query-modality alignment of embeddings



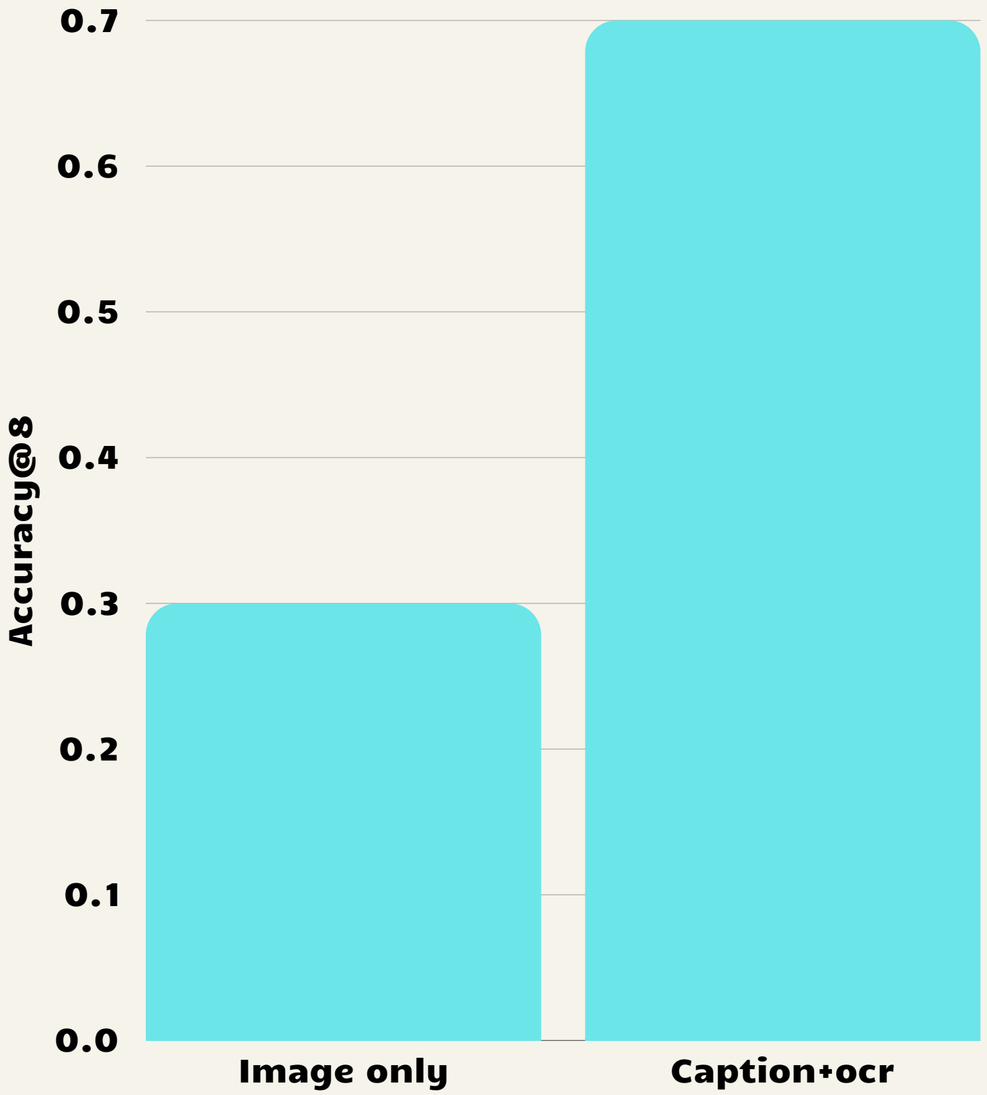
Metadata Improves Image Retrieval

Impact of Metadata on Image Retrieval

Accuracy@8 on image-based queries:

- Image embeddings: **0.30**
- Metadata(Caption + OCR) embeddings: **0.70**

- Using only raw CLIP image embeddings the retriever performs poorly: 30% accuracy.
- Using metadata (caption + ocr text) the accuracy improve to 70%.
- This guided our final design choice: retrieval via metadata embedding only.



Sample queries and responses

Query: In the bar chart that compares the train and test errors for all evaluated ML algorithms, which two models showed the highest test error

Response: Based on the bar chart in Figure 4 and the precise values given in Table 2, the two models that showed the highest test error are:

1. **K-Nearest Regression:** Mean test error of **0.54**
2. **Linear Regression** and **Ridge Regression:** Both had a mean test error of **0.53**

Query: On the volcano plot showing the Turnover Frequency for acetaldehyde production from NODH of ethanol, are the ML-predicted TOFs for Cu_3Ni and Cu_3Pt alloys higher or lower than their DFT-calculated counterparts

Response: Based on the provided text, the following can be inferred:

1. **Volcano plots for acetaldehyde production from NODH of ethanol** are mentioned (Fig. 14(a) and (b)). These plots use "descriptor energies obtained from DFT modelling."
2. **ML is used to predict binding energies** for alloys, which are then used in microkinetic modeling (MKM) to derive reaction energetics.
3. The text does make a comparison between DFT and ML calculated **alloy energetics** (binding energies, not directly TOFs) for a list of alloys.

However, the text **does not provide enough specific information to directly answer your question** about whether ML-predicted TOFs for Cu_3Ni and Cu_3Pt alloys are higher or lower than their DFT-calculated counterparts on the volcano plot:

- **For Cu_3Ni (mentioned as Ni_3Cu in the text):** The text compares DFT and ML calculated *alloy energetics* for `Ni3Cu`. It shows a difference in the ordering of `Ni3Cu` relative to other alloys when comparing DFT and ML energetics:
  - DFT energetics: `... > Ni3Cu > Ni3Pt > Ni3Rh`
  - ML energetics: `... > Ni3Cu Ni3Pt Ni3Rh` While this indicates a difference in the underlying binding energies, the text **does not explicitly state how this difference translates into higher or lower Turnover Frequencies (TOFs) on the volcano plot** for `Ni3Cu`. The volcano plot itself, as described, is based on DFT-derived descriptor energies.
- **For Cu_3Pt:** The provided text **does not mention Cu_3Pt (or Pt_3Cu) at all**. It mentions `Ni3Pt`, but this is a different alloy. Therefore, no information is available to answer for Cu_3Pt.

In summary, the text indicates that ML and DFT derive different alloy energetics, which would lead to different predicted TOFs if both were plotted. However, it **does not explicitly show the volcano plots comparing both DFT- and ML-predicted TOFs for Cu_3Ni or Cu_3Pt, nor does it state which would be higher or lower for these specific alloys.**

Conclusion

- Developed a multimodal RAG assistant capable of extracting and synthesizing **text, tables, and scientific figures** from catalysis PDFs
- Achieved robust retrieval performance: **85%** for text-based queries and **70%** for image-based queries
- **Metadata-driven retrieval** proved essential — captions + OCR significantly improved image retrieval accuracy
- System successfully answers cross-modal scientific questions, demonstrating practical usefulness for catalyst research
- Validates the **feasibility of automated literature mining** to reduce manual data extraction and accelerate catalyst discovery

FUTURE WORK

1. **Improve Figure–Caption Alignment** (Highest Priority, High Feasibility)
    - **Improves**: Retrieval accuracy for scientific images
    - **How**: Use layout-aware models (LayoutLMv3 / DocFormer) to correctly bind figures to their captions, reducing caption mismatch and multi-figure confusion.
2. **Advanced Scientific Figure Understanding** (Medium Priority, Medium Feasibility)
    - **Improves**: Interpretation of complex plots (volcano, heatmaps, kinetics graphs)
    - **How**: Integrate chart-parsing and digitization tools (ChartOCR, UniChart) to extract axes labels, data points, and structured values — enabling deeper reasoning, not just retrieval.
3. **Learned Multimodal Fusion & Re-ranking** (High Impact, Lower Feasibility)
    - Improves: Consistency and accuracy of multimodal search
    - How: Train fusion layers or cross-encoders that jointly embed (query + image metadata) and re-rank top-k results, allowing more precise alignment than raw CLIP embeddings.

THANK YOU

For your attention