香港中文大學
The Chinese University of Hong Kong

Materials Synthesis Insights from Scientific Literature via Text Extraction and Machine Learning

Isha Kumari

20 June 2025

Chem. Mater. 2017, 29, 9436-9444

Contents

- Background
- Methodology
- Results
- Conclusion
- Future Outlook

Background

Background

Why This Matters: The Need for Data-Driven Synthesis

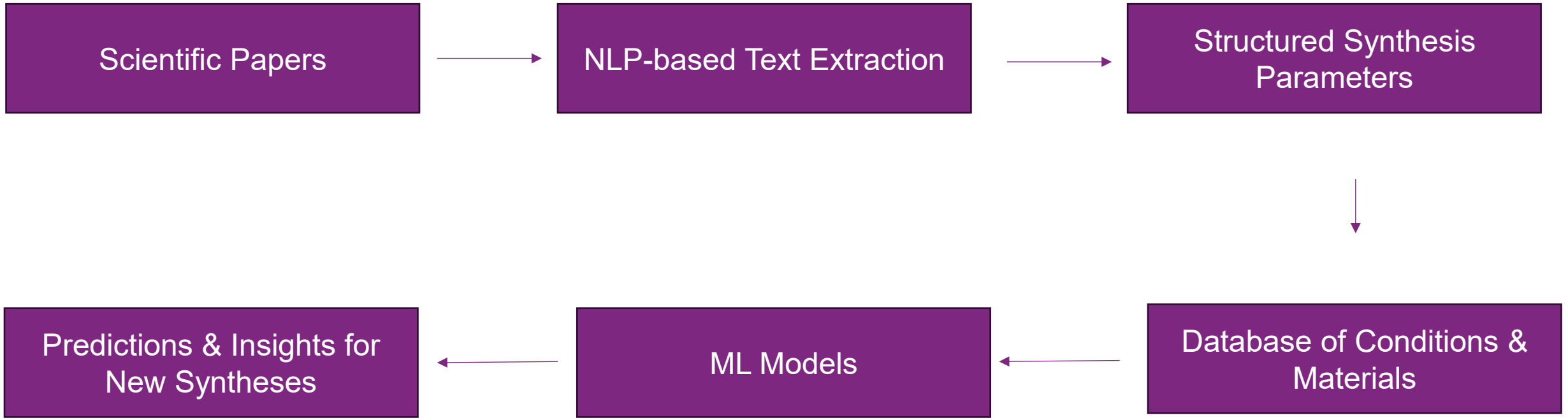
- The pace of discovering new materials is often slowed by the lack of efficient synthesis strategies.
- Traditionally, synthesis relies on heuristics, which do not scale with the rapid growth of computational material predictions
- This project leverages **Natural Language Processing (NLP)** and **Machine Learning (ML)** to extract and analyze synthesis parameters from thousands of published scientific articles.

By building a large, structured database of synthesis conditions, we can:

- Understand key parameters influencing material properties
- Improve reproducibility and guide new syntheses
- Enable data-driven automation in materials design

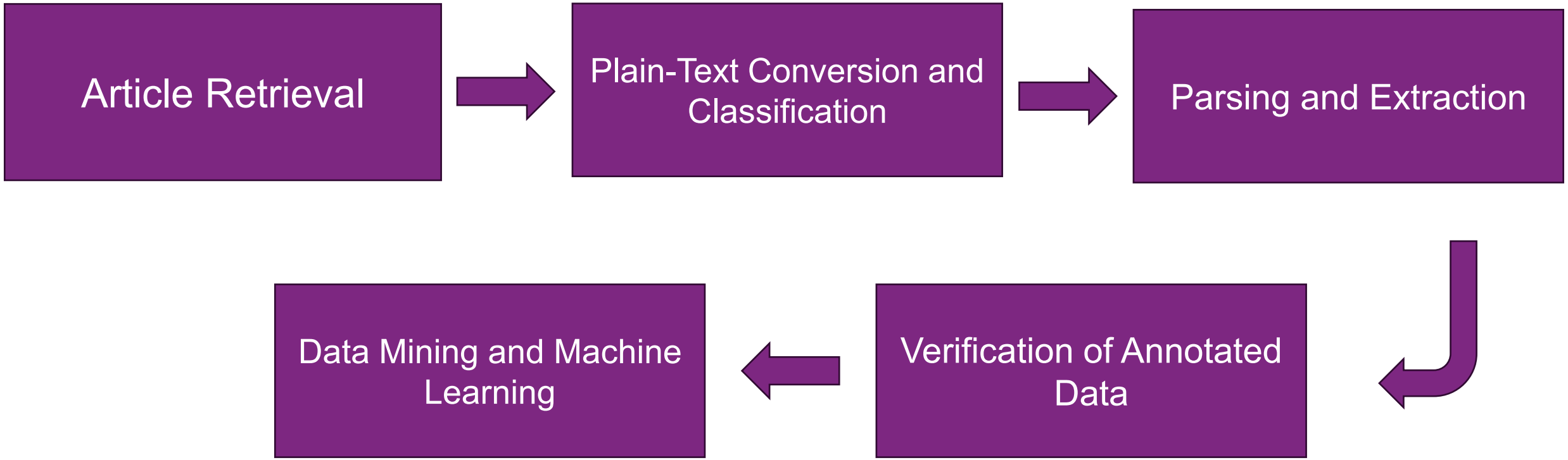
Background

Framework for Literature-Based Synthesis Extraction

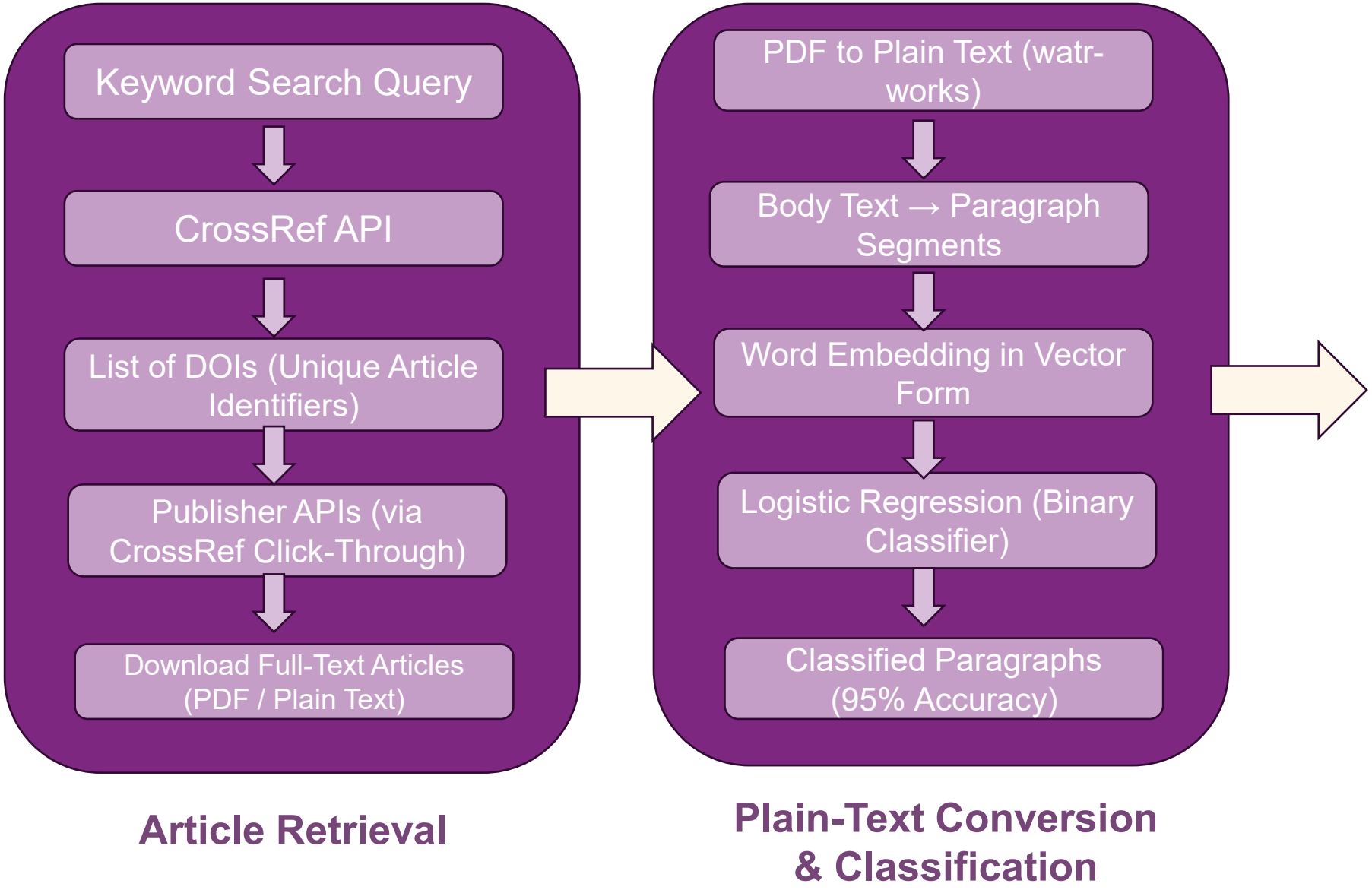


Methodology

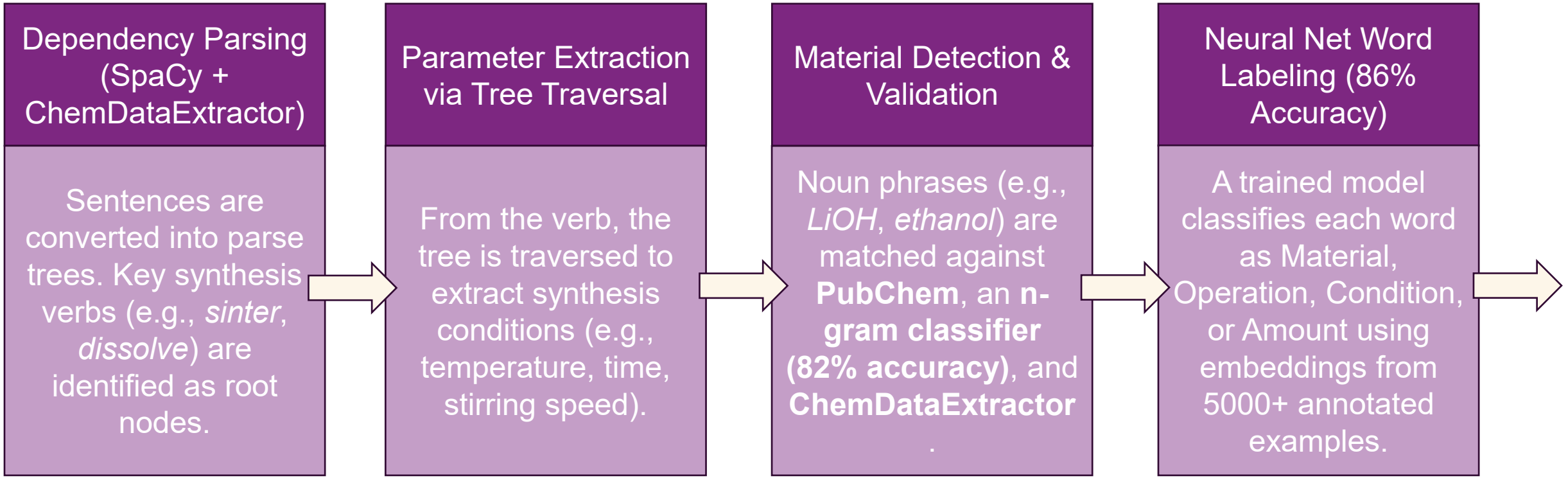
Methodology



Methodology

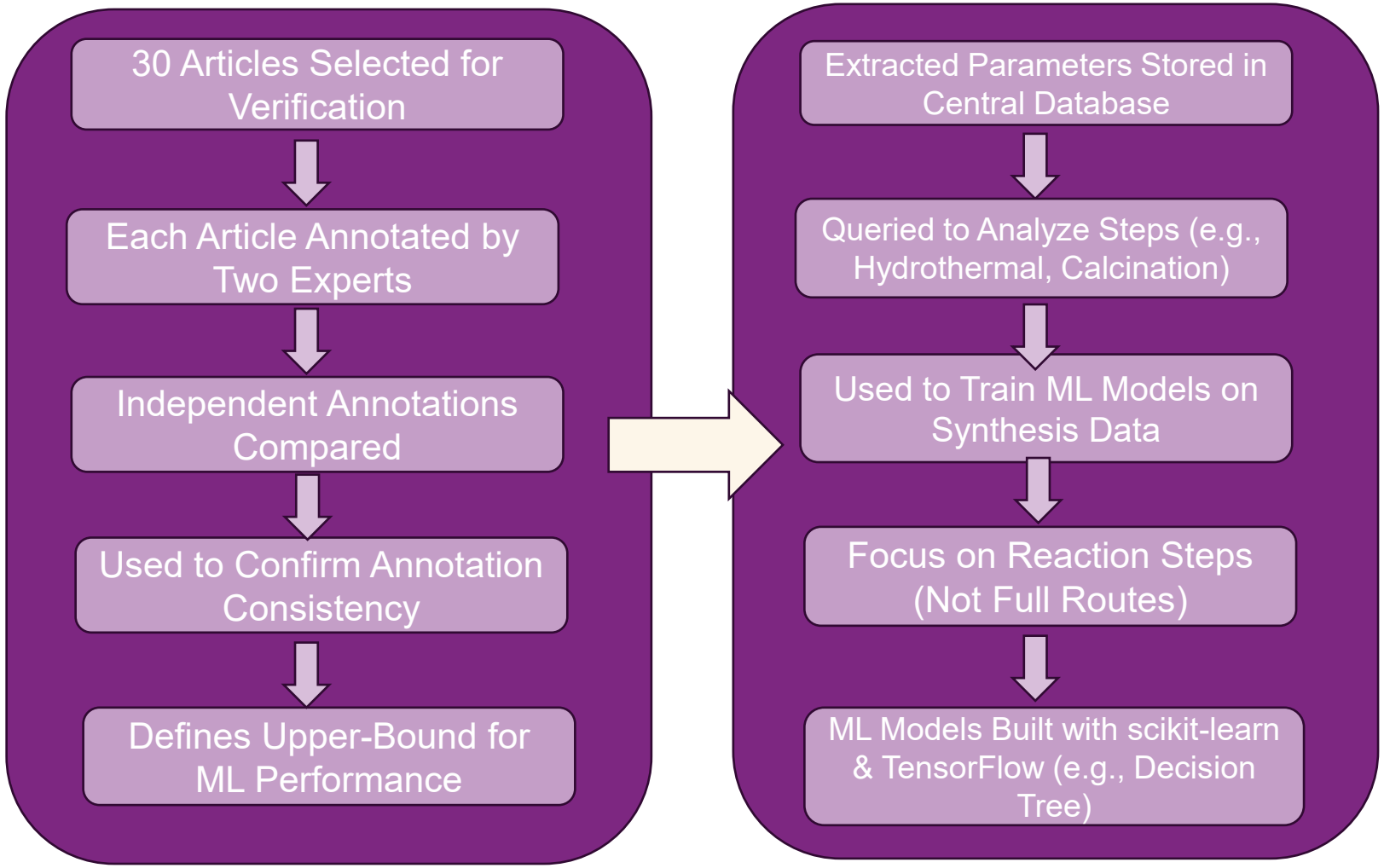


Methodology



Parsing and Extraction

Methodology

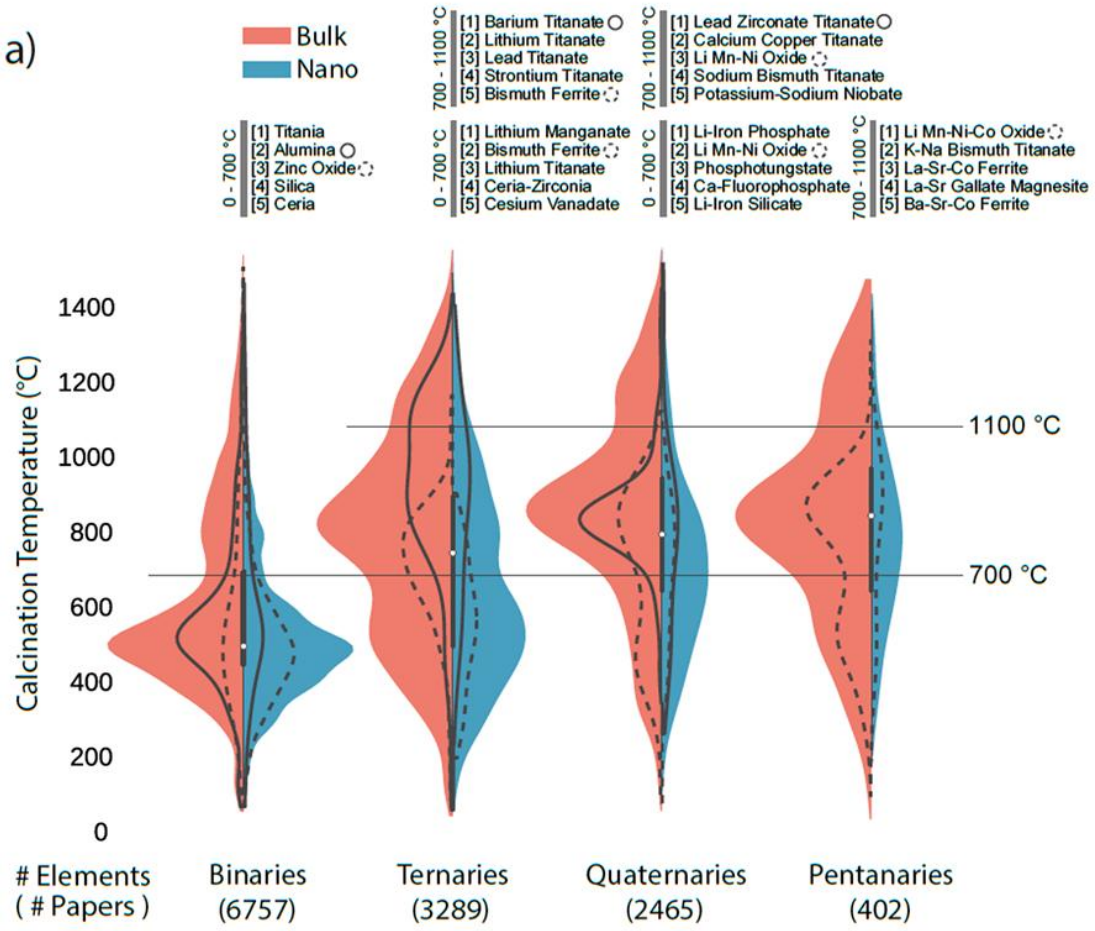


Verification of Annotated Data

Data Mining and Machine Learning

Results

Results



Calcination Temperature Distribution Across Metal Oxide Systems (Bulk vs Nano)

Bulk vs Nano Trend:
Bulk materials are typically calcined at **higher temperatures** than their nanostructured counterparts.

Complexity Increases Temperature:
Calcination temperature increases with **elemental complexity,** from binaries to pentanaries, especially in **nano** syntheses.

Binary Oxide Range:
Most binary oxides fall between **450–550 ° C**

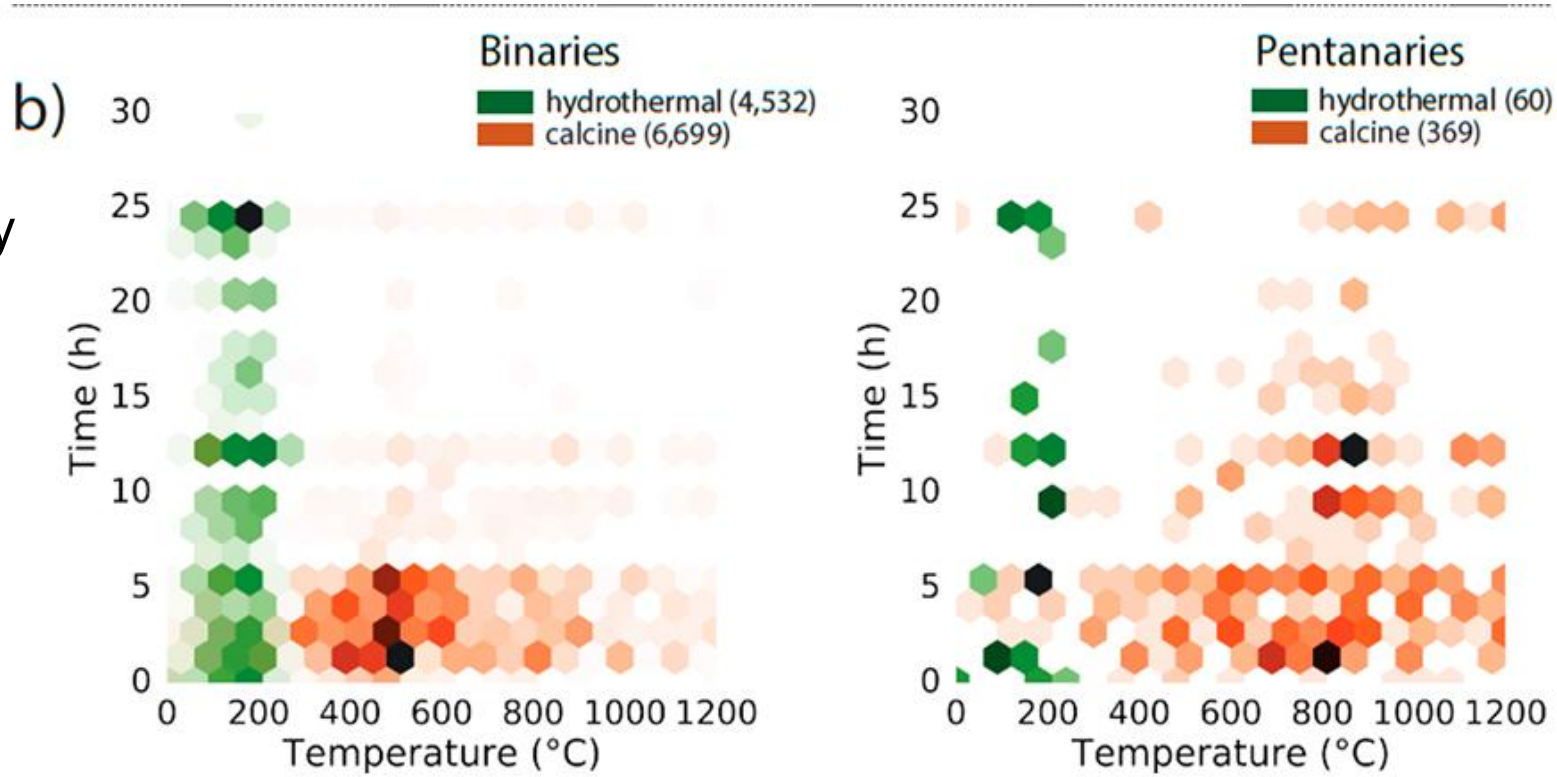
- **Alumina:** found more in bulk
- **ZnO:** more in nano form

Ternary oxide Range:

- **Bismuth Ferrite:** sharp peak at 750 ° c(bulk) and around 600 ° C for nano.
- **Barium Titanate:** wider range, 900-1100 ° C

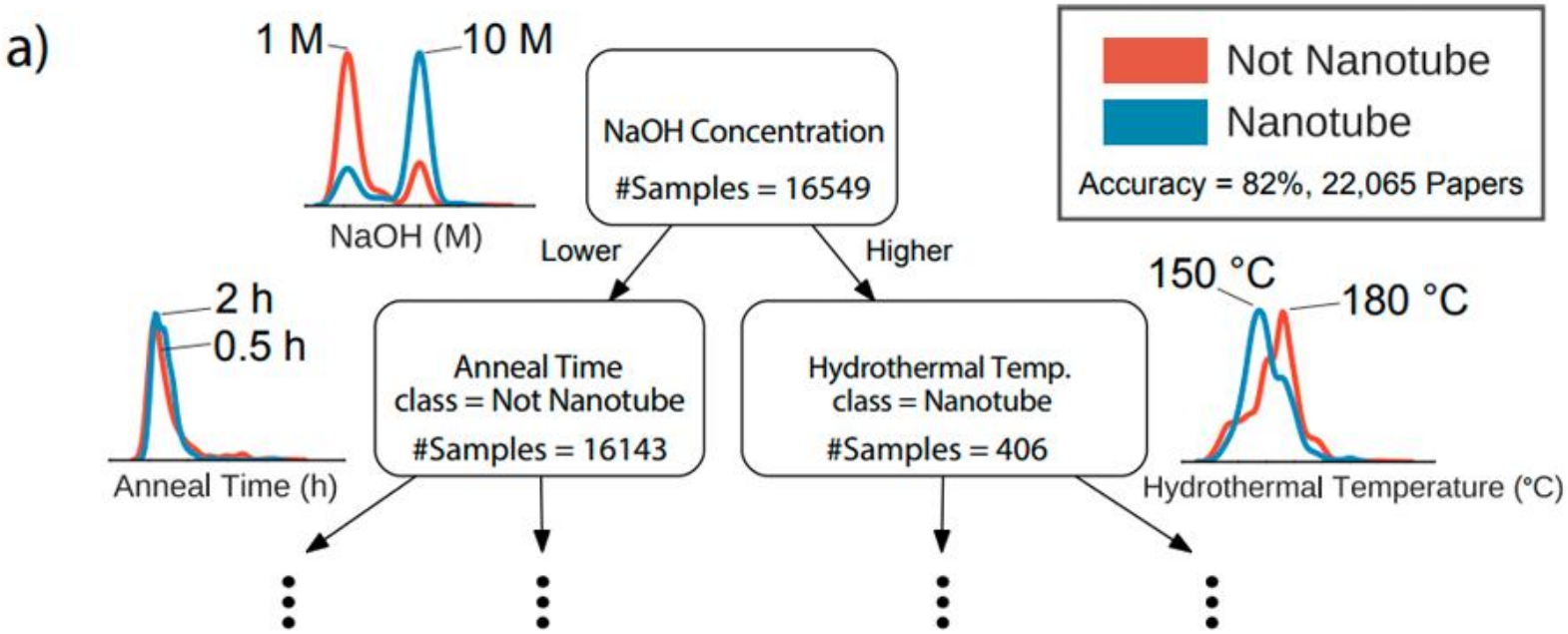
Results

- **Hydrothermal reactions** occur at **low temperatures** (typically 150–200 °C) and **long durations** (12–24 hours).
- Reaction temperatures are limited by the **critical point of common solvents** (e.g., water, ethanol).
- **Calcination steps** happen at **much higher temperatures** and usually **shorter times**.
- As material complexity increases, **calcination temperatures and durations both rise**.
- Simple systems concentrate around **400–500 °C, <5 h**, while complex ones extend to **800–900 °C, >5 h**.



Temperature–Time Distribution of Hydrothermal and Calcination Syntheses

Results



Decision Tree for Titania Nanotube Formation

Key Features Identified:

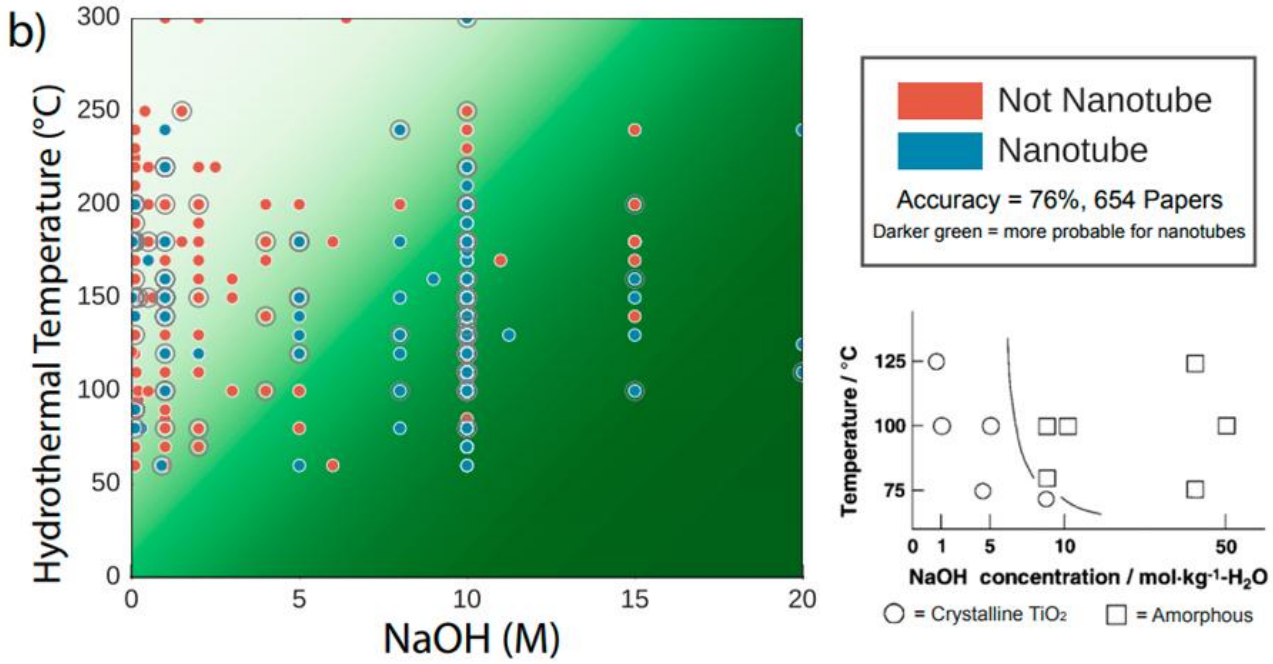
NaOH concentration (most important , root node)
is commonly clustered around **1 M and 10 M**

Hydrothermal temperature also plays a significant role
Some notable peaks at **150° C and 180° C**

Annealing time was found to be **less predictive**, as its values didn't clearly distinguish between outcomes.

The model achieves 82% classification accuracy, showing that NaOH concentration and hydrothermal temperature are strong indicators of nanotube formation.

Results



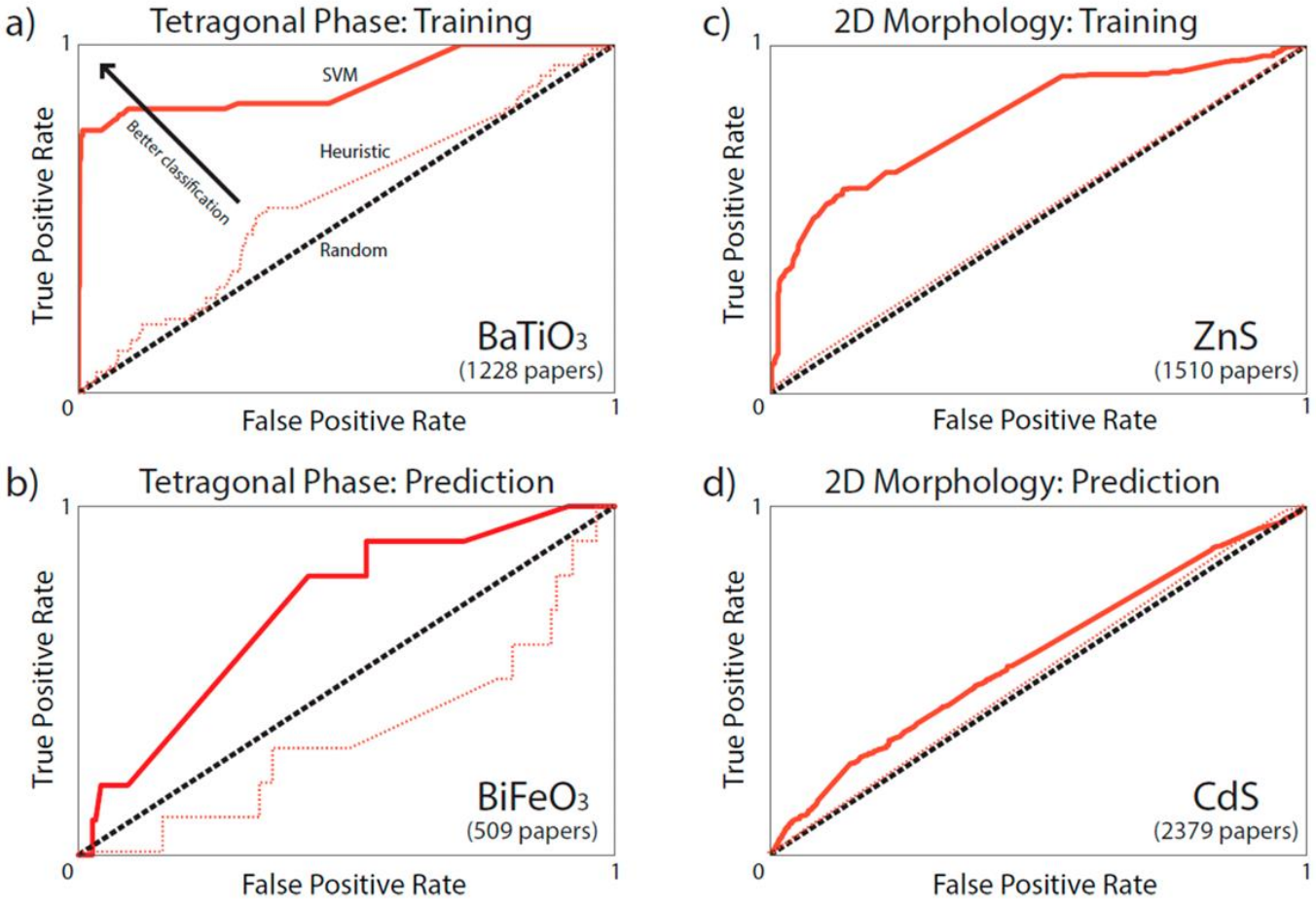
Machine-Learned Phase Diagram for Titania Nanotube Formation

The diagram shows higher probability of nanotube formation (darker green) at:

- High NaOH concentrations
- Lower hydrothermal temperatures

This reduced parameter space allows for easy visualization of synthesis trends using only experimentally accessible variables.

Results



Machine-learned classifiers and predictions across materials systems

All four subfigures compare three models:

– **Support Vector Machine (SVM)** with Gaussian kernel

– **Heuristic (simple rule-based)**

– **Random guessing**

SVM consistently outperforms both the heuristic and random strategies.

Demonstrate **robustness and transferability of machine learning models** across materials and synthesis outcomes.

Conclusion

Conclusion

This work bridges the gap between published knowledge and predictive synthesis planning.

Built an **automated framework** to extract synthesis data using **text mining and ML**.

Compiled a **large-scale database** of synthesis conditions from scientific literature.

Identified **key parameters** (e.g., temp, time, solvent) that influence material outcomes.

Enabled **accurate predictions** without prior domain knowledge or manual input.

Future Outlook

Future Outlook

Extend coverage to more synthesis methods

Thin film deposition, catalytic systems, non-crystalline materials

Refine data accuracy

Weighting **highly cited papers**
Incorporating **newly published research**

Thank You

Any Questions?