Automating Materials Science Research: NLP-Driven Pipeline for Data Analysis and Experimentation

Zhang Yuchen[a,b] , **Chen Bin**[b] , **Su Jian**[b]

[a] *CNRS@CREATE LTD, Singapore*

[b] *Institute for Infocomm Research (I2R), A*STAR, Singapore* {zhangyuc, bchen, sujian}@i2r.a-star.edu.sg

1. Introduction

The rapid expansion of scientific data— encompassing experiment data stored in databases, reported in the literature, or used for computational simulations —has created significant opportunities and challenges for researchers across various areas. Efficiently extracting, analyzing, and leveraging this expanding data is essential for accelerating innovation and optimizing workflows. However, the complexity of database queries and domain-specific computational analysis often presents significant barriers, especially for researchers lacking expertise in database management or programming.

This paper explores how recent advancements in natural language processing (NLP) and generative AI, specifically Text-to-SQL(T2S)[1] and its more general form Text-to-Code(T2C)[2] technologies, can address these challenges, focusing on applications within materials science. We present an integrated pipeline that streamlines material science data retrieval, analysis, and simulation. Specifically, the pipeline utilizes T2S to query structured relational databases (e.g., established repositories like Materials Project[3], AFLOW[4], and local repositories) and T2C to generate scripts for extracting data from semi-structured sources (e.g., JSON files, experimental logs). This combination offers a flexible and comprehensive solution for accessing diverse data resources. Furthermore, building on the retrieved data, the pipeline enables automated data analysis with expert-defined functions and supports simulation setup with high-throughput analysis via Text–to–Code.

As the established databases cover many topics and often lag behind the most recent published work, information extraction(IE) is incorporated to extract experimental related information automatically from the most recent publications / online information release. The extracted information is stored in local databases for downstream analysis through T2S.

This work aims to significantly enhance the efficiency, scalability, and accessibility of material science research, ultimately accelerating innovation and discovery. The same techniques and framework could be easily adapted to other scientific domains, such as biomedical research, chemical engineering, pharmaceutical development, and climate / environmental science, where large-scale literature analysis and structured data extraction are equally valuable.

2. Related work

Code generation has been widely adopted across various scientific disciplines to automate repetitive tasks, optimize workflows, and enhance reproducibility. For example, in bioinformatics, tools like Galaxy[5] and BioPython[6] have enabled researchers to generate code for complex genomic data analysis pipelines, reducing the need for manual coding and minimizing errors. In physics, frameworks such as FiPy[7] have been used to generate code for symbolic mathematics and partial differential equation solvers. These examples demonstrate the potential of code generation to streamline workflows and improve efficiency across various scientific domains. Despite the growing complexity of computational models and the increasing volume of experimental data, T2C adoption in materials science remains limited[8], highlighting an opportunity for innovation. The SOTA T2C, achieving 87.2% accuracy in the SemEval-2025 benchmark on the Tabular QA application[9], demonstrates its readiness.

SOTA T2S systems like DuoSQL[10] and DAIL-SQL combined with large language models (LLMs)[11] have effectively simplified data retrieval in general-purpose databases. Extending these techniques to materials science databases could significantly enhance the researcher's efficiency to exploit these repositories with vast amounts of structured materials data. We intend to build on our work using a generative approach for T2S, which leveraging LLMs with a chain-of-verification procedure, has demonstrated the ability to generate high-quality, large-scale SQL datasets at a controlled cost, effectively addressing complex real-world queries.[12].

IE plays a crucial role in automatically converting the unstructured text into the structured information stored in to the database for downstream analysis. There're limited studies on material science IE primarily focusing on extractions of named entities such as materials and material properties as well as entity relations[13] such as capacity and conductivity. The first event extraction dataset for material science is SC-CoMIcs[14] on doping events in superconductivity. Leverage on SC-CoMIcs, we explore hybrid discriminative as well as generative model to further advance the technologies to serve the purpose here and support the knowledge discovery in general[15].

With these advanced NLP technologies, we propose a pipeline that automates data retrieval / anal-

ysis and supports simulation setup / analysis, ultimately accelerating materials discovery.

3. A NLP Pipeline for Materials Science Research

Figure 1 illustrates the overall workflow of our system. The process begins with a user-submitted query processed through our pipeline. The first stage, Data Query, employs T2S techniques to retrieve relevant information from structured databases (remote and local) and T2C techniques to extract data from semi-structured sources. This step yields an initial dataset, serving as the foundation for further analysis.

Subsequently, in the User-Defined Analysis stage, we apply T2C techniques to process the initial dataset according to user specifications. This could include data visualization, complex analysis, or simulation setup tasks, enabling customized analytical workflows tailored to specific research or application needs.

The information extraction modules work at the backend on extracting experimental data from texts and tables from literature.

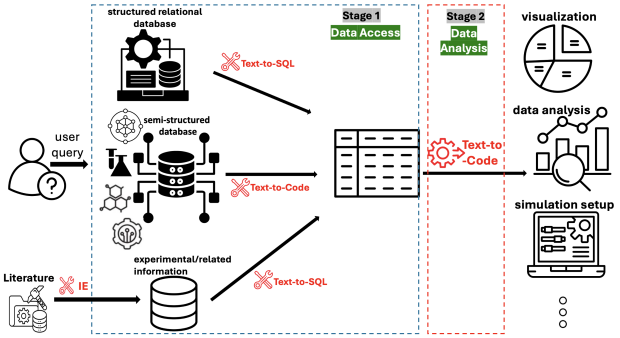

Fig. 1: NLP-Driven Pipeline Workflow

3.1 T2S for Database Access

Material properties, synthesis parameters, and performance metrics are often stored in relational databases. T2S technologies significantly simplify the data retrieval process from prominent materials databases like MatWeb[16], Materials Project, AFLOW, and the focused local databases, enabling intuitive access for researchers without SQL expertise. For example:

Natural Language Query: *"Screen organic photovoltaic materials in a combinatorial library with a power conversion efficiency (PCE) > 12%, open-circuit voltage (Voc) > 0.8V, and a scalable synthesis score in the top quartile."*

Figure 2 shows the generated SQL and query output. T2S is used to query the material library database based on multi-dimensional performance criteria. The system generates a query that selects materials meeting all required conditions and excludes candidates not meeting scalability or performance benchmarks.

```sql
SELECT material_id, PCE, Voc, synthesis_score
FROM organic_photovoltaics
WHERE PCE > 12
  AND Voc > 0.8
  AND synthesis_score >=
      (SELECT percentile_score
       synthesis_scores WHERE percentile = 75);
```

output:

material_id	PCE	Voc	synthesis score
OPV_001	13.5	0.85	80
OPV_004	14.2	0.88	90

Fig. 2: SQL Query from Structured Database

3.2 T2C for Simulations, Visualization and Analysis

Many materials properties are determined through computational methods like Density Functional Theory (DFT) and Molecular Dynamics (MD). T2C tools can generate Python scripts for simulations based on user queries. For instance:

```python
from ase import Atoms
from ase.calculators.vasp import Vasp

mol = Atoms('MoS2', positions=[[0,0,0], [1.5,1.5,1.5]])
calc = Vasp(xc='PBE', encut=400, kpts=[3,3,3])
mol.set_calculator(calc)
energy = mol.get_potential_energy()
print(f"Total energy: {energy} eV")
```

output:

```
Total energy: -500.123 eV
```

Fig. 3: calculate the total energy using DFT and VASP

Figure 3 shows that T2C can automatically generate Python scripts that interface with tools like VASP for material simulations. Natural language descriptions are converted into Python code that sets up a molecular system, applies the VASP calculator with specified parameters (e.g., PBE functional, ENCUT, kpts), and computes the system's potential energy. This automation streamlines material modeling and energy calculations, speeding up material discovery.

Leveraging T2C technologies, researchers can also intuitively execute complex visualization and data analysis tasks directly from natural language queries, significantly reducing the requirement for programming expertise.

4. Conclusion

We propose an NLP pipeline incorporating T2S, T2C, and IE to enable materials science researchers to easily access experiment-related information stored in structured, semi-structured, and unstructured formats using natural language. The extracted data can be further visualized, analyzed, and simulated for seamless knowledge discovery. We briefly discuss the ongoing and future extensions of these NLP technologies in the field of materials science.

These technical advancements collectively enhance our proposed pipeline's capability, positioning us strongly for future expansion in automation, robustness, and cross-disciplinary applicability

Acknowledgments

This research is part of the programme DesCartes and is supported by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme.

References

[1] Zijin Hong, Zheng Yuan, Qinggang Zhang, Hao Chen, Junnan Dong, Feiran Huang, and Xiao Huang. Next-generation database interfaces: A survey of llm-based text-to-sql, 2025.

[2] Juyong Jiang, Fan Wang, Jiasi Shen, Sungju Kim, and Sunghun Kim. A survey on large language models for code generation, 2024.

[3] Materials Project. Next-generation materials project, 2025. Accessed: 2025-03-09.

[4] Camilo E. Calderon, Jose J. Plata, Cormac Toher, Corey Oses, Ohad Levy, Marco Fornari, Amir Natan, Michael J. Mehl, Gus Hart, Marco Buongiorno Nardelli, and Stefano Curtarolo. The aflow standard for high-throughput materials science calculations, 2015.

[5] Daniel Blankenberg, Assaf Gordon, Gregory Von Kuster, Nathan Coraor, James Taylor, Anton Nekrutenko, and the Galaxy Team. Manipulation of fastq data with galaxy. *Bioinformatics*, 26(14):1783–1785, 06 2010.

[6] BioPython Contributors. Biopython: Open-source tools for computational biology, https://biopython.org/.

[7] Jonathan E. Guyer, Daniel Wheeler, and James A. Warren. Fipy: Partial differential equations with python. *Computing in Science Engineering*, 11(3):6–15, 2009.

[8] Santiago Miret and Nandan M Krishnan. Are llms ready for real-world materials discovery? *arXiv preprint arXiv:2402.05200*, 2024.

[9] Bin Chen Gao Yuze and J. Su. I2r-nlp at semeval-2025 task 8: Question answering on tabluar data. Submitted to SemEval-2025, 2025.

[10] Weikang zhang, Zhi Liu, Tongxin Bai, Furong Zheng, Wenming Jin, and Yang Wang. Duosql: towards elastic data warehousing via separated data management and processing. *The Computer Journal*, page bxaf014, 02 2025.

[11] Dawei Gao, Haibin Wang, Yaliang Li, Xiuyu Sun, Yichen Qian, Bolin Ding, and Jingren Zhou. Text-to-sql empowered by large language models: A benchmark evaluation, 2023.

[12] Yuchen Zhang, Yuze Gao, Bin Chen, Wenfeng Li, Shuo Sun, and Jian Su. High-quality complex text-to-sql data generation through chain-of-verification. Submitted to ACL 2025, 2025.

[13] Yu Song, Santiago Miret, and Bang Liu. Matscinlp: Evaluating scientific language models on materials science language tasks using text-to-schema modeling, 2023. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023.

[14] Kyosuke Yamaguchi, Ryoji Asahi, and Yutaka Sasaki. SC-CoMIcs: A superconductivity corpus for materials informatics. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6753–6760, Marseille, France, May 2020. European Language Resources Association.

[15] A. Hao and J. Su. Hybrid and generative models for material science event extraction. Submitted to AI4X 2024, Singapore, August 8-11, 2025, 2025.

[16] MatWeb. Matweb materials database, https://www.matweb.com/.