CIERC — Collective Intelligence for the Exploration of the Reticular Chemistry Synthesis Space

Dongrong Joe Fu[*1] Nakul Rampal[*2,3,4] Ruikang Wang[1] Zihui Zhou[2,3,4] Christian Borgs[1,4*] Omar M. Yaghi[2,3,4,5*] Jennifer T. Chayes[1,4,6,7,8*]

[*]Equal contribution [1]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, California 94720, USA [2]Department of Chemistry, University of California, Berkeley, California 94720, USA. [3]Kavli Energy Nanoscience Institute, University of California, Berkeley, California 94720, USA [4]Bakar Institute of Digital Materials for the Planet, College of Computing, Data Science, and Society, University of California, Berkeley, California 94720, USA. [5]KACST-UC Berkeley Center of Excellence for Nanomaterials for Clean Energy Applications, King Abdulaziz City for Science and Technology, Riyadh 11442, Saudi Arabia [6]Department of Mathematics, University of California, Berkeley, California 94720, USA [7]Department of Statistics, University of California, Berkeley, California 94720, USA [8]School of Information, University of California, Berkeley, California 94720, USA

Correspondence to: borgs@berkeley.edu yaghi@berkeley.edu jchayes@berkeley.edu

1. Introduction

Reticular chemistry offers extraordinary flexibility in material design. For instance, a wide range of metal nodes and organic linkers can, in principle, form MOFs. However, this flexibility also creates a fundamental challenge: synthesis outcomes depend on many interacting variables whose relationships to structure, stability, and performance are partially unrevealed. As a result, synthesis conditions often function as a black box, guided largely by empirical precedent rather than a sound and predictive theory. Despite decades of experimental progress, the field lacks a comprehensive, machine-readable synthesis database. Instead, critical synthesis details are scattered across the literature in heterogeneous formats, including narrative text, tables, figures, and crystallographic files, and are frequently split across main manuscripts (MS) and supplementary information (SI). This fragmentation prevents systematic reuse of prior knowledge and limits the applicability of data-driven methods. Recent advances in LLMs suggest new opportunities for scientific knowledge extraction and reasoning. However, applying LLMs to synthesis understanding requires addressing both the unstructured nature of the literature and the domain-specific complexity of chemical synthesis. This motivates AI framework that aim to make empirical synthesis knowledge explicit and usable for reasoning, rather than treating synthesis as a purely predictive task.

In this work, we introduce CIERC (Collective Intelligence for the Exploration of the Reticular Chemistry Synthesis Space), a domain-aware AI framework for organizing and reasoning over fragmented synthesis knowledge. CIERC converts unstructured literature into synthesis records, adapts large language models through domain pretraining and task alignment, and uses embedding-based specialization for localized synthesis reasoning and confidence estimation. Rather than assuming predictive theory, CIERC leverages empirical precedent across the literature.

2. Methods

CIERC follows a multi-stage pipeline designed to transform fragmented synthesis literature into a domain-aware AI framework for reticular chemistry synthesis reasoning. The pipeline explicitly separates text-based knowledge extraction, model adaptation, and synthesis-aware specialization, reflecting both the structure of the literature and the complexity of the reticular chemistry synthesis space.

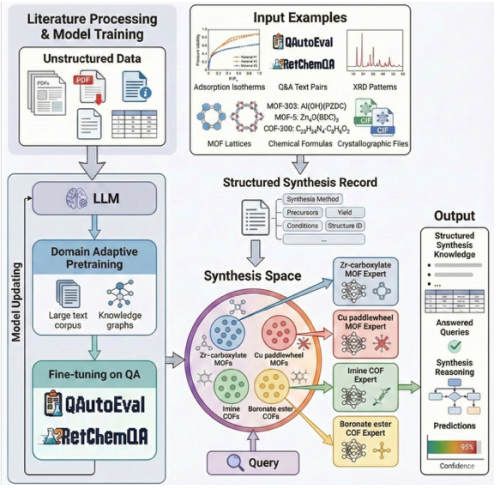

Fig. 1: Overview of the CIERC pipeline for reticular chemistry synthesis reasoning.

2.1 Text-based literature processing.

Initial stages use raw text from main manuscripts and supplementary information, including synthesis descriptions, formulas, reaction parameters, and contextual

statements. Tables, crystallographic files, and other non-text data are excluded. The goal is to preserve experimental narrative while converting text into structured representations that identify synthesis passages, resolve MS–SI links, and maintain protocol continuity.

2.2 Domain-adaptive model training.

The structured text corpus derived from MS and SI is first used for continual domain-adaptive pretraining (DAPT) of Qwen3-14B-instruct. This step adapts the model to the linguistic patterns, terminology, and conceptual structure of reticular chemistry synthesis literature, providing domain fluency without task-specific bias.

Following DAPT, we perform a first round of supervised fine-tuning to align the model with synthesis-centric tasks. This stage focuses on stabilizing domain-specific reasoning and grounding the model in synthesis-relevant behaviours, such as identifying key synthesis variables, comparing reported protocols, and answering synthesis-related questions based strictly on textual evidence. Importantly, this fine-tuning stage relies exclusively on text-derived data and does not incorporate curated databases or non-text experimental artifacts, ensuring that the model's behaviour reflects reported experimental knowledge rather than post hoc structure annotations.

2.3 Embedding-based synthesis record construction and specialization.

Building on our previous work (RetChemQA) [1], which focuses on text-based synthesis condition extraction and synthesis-related question–answering, we extend the framework to incorporate new representations of non-textual sources, including crystallographic files, tabulated parameters, and other structured experimental artifacts, to construct embedded synthesis records. In a later stage of the pipeline, these data are used to enrich synthesis representations with structural and experimental context. The resulting synthesis records are embedded into a shared representation space, where unsupervised clustering reveals natural reticular chemistry domains corresponding to distinct experimental patterns.

Lightweight expert models are trained for each cluster in the reticular chemistry synthesis space to capture localized empirical knowledge. During inference, synthesis queries are embedded and routed to the most relevant experts based on similarity in this representation space. Embedding distance additionally serves as a confidence signal, reflecting proximity to existing synthesis precedent and distinguishing well-supported synthesis queries from exploratory conditions.

By separating domain adaptation, task alignment, and synthesis specialization, CIERC enables AI-assisted reasoning over synthesis conditions without assuming predictive theory, instead organizing empirical synthesis knowledge across the literature. This allows CIERC to distinguish well-supported synthesis queries from exploratory regimes based on empirical support.

3. Related work

LLMs have recently been applied to chemistry literature mining, including the extraction of MOF synthesis conditions from unstructured experimental text [2]. Recent work has also explored literature-informed knowledge integration for MOF research, such as MOF-ChemUnity, which links literature knowledge, crystal structures, and computational datasets into a unified knowledge graph supporting literature-informed AI reasoning [3]. Other frameworks integrate LLMs with curated databases and predictive models to support materials discovery and property prediction [4]. These approaches primarily focus on parameter extraction or downstream prediction, and do not address how fragmented synthesis knowledge can be systematically organized to support synthesis reasoning.

Our embedding-based synthesis record construction and specialization strategy (Section 3.3) is inspired by recent work on collective intelligence for AI-assisted chemical synthesis, which demonstrates that partitioning chemical space into specialized experts improves reliability and interpretability [5]. We adapt this paradigm to reticular chemistry by organizing literature-derived synthesis records into synthesis condition space aware representations that support reasoning and confidence estimation, rather than direct protocol generation. Unlike prior frameworks that focus on parameter extraction or forward prediction, CIERC is designed to organize fragmented synthesis knowledge into a structure that supports synthesis reasoning and uncertainty awareness.

Acknowledgments

N. R. and Z. Z. acknowledge the Bakar Institute of Digital Materials for the Planet (BIDMaP) Emerging Scholars Program for the funding that supports this work.

References

[1] Nakul Rampal, Kaiyu Wang, Matthew Burigana, Lingxiang Hou, Juri Al-Johani, Anna Sackmann, Hanan S. Murayshid, Walaa A. AlSumari, Arwa M. Al-Abdulkarim, Nahla E. Alhazmi, Majed O. Alawad, Christian Borgs, Jennifer T. Chayes & Omar M. Yaghi. Single and Multi-Hop Question-Answering Datasets for Reticular Chemistry with GPT-4-Turbo. Journal of Chemical Theory and Computation 20 (20), 9128–9137 (2024). DOI: 10.1021/acs.jctc.4c00805.

[2] Zhiling Zheng, Oufan Zhang, Christian Borgs, Jennifer T. Chayes, and Omar M. Yaghi. ChatGPT Chemistry Assistant for Text Mining and the Prediction of MOF Synthesis. Journal of the American Chemical Society, 145(33):18048–18062, 2023.

[3] Thomas Michael Pruyn et al. MOF-ChemUnity: Literature-Informed Large Language Models for Metal–Organic Framework Research. JACS, 2025.

[4] Yeonghun Kang and Jihan Kim. ChatMOF: An artificial intelligence system for predicting and generating metal–organic frameworks using large language models. Nature Communications, 15:4705, 2024.

[5] Haote Li, Sumon Sarkar, Wenxin Lu, Patrick O. Loftus, Tianyin Qiu, Yu Shee, Abbigayle E. Cuomo, … & Victor S. Batista. Collective intelligence for AI-assisted chemical synthesis. Nature (2026). DOI: 10.1038/s41586-026-10131-4.