Leveraging Chemistry Foundation Models to Facilitate Structure Focused Retrieval Augmented Generation in Multi-Agent Workflows for Catalyst and Materials Design

Nathaniel H. Park
IBM Research—Almaden
San Jose, CA 95120
npark@us.ibm.com
Tiffany Callahan
IBM Research–Almaden
San Jose, CA 95120
James L. Hedrick
IBM Research—Almaden
San Jose, CA 95120
hedrick@us.ibm.com
Tim Erdmann
IBM Research—Almaden
San Jose, CA 95120
Tim.Erdmann@ibm.com
Sara Capponi
IBM Research—Almaden
San Jose, CA 95120
Sara.Capponi@ibm.com

Abstract

Molecular property prediction and generative design via deep learning models has been the subject of intense research given its potential to accelerate development of new, high-performance materials. More recently, these workflows have been significantly augmented with large language models (LLMs) and systems of autonomous agents capable of utilizing pre-trained models to make predictions in the context of more complex research tasks. While effective, there is still room for substantial improvement within agentic systems on the retrieval of salient information for material design tasks. Within this context, alternative uses of deep learning models, such as leveraging their latent representations to facilitate cross-modal retrieval augmented generation within agentic systems for task-specific materials design, has remained unexplored. Herein, we demonstrate that large, pre-trained chemistry foundation models can serve as a basis for enabling structure-focused, semantic chemistry information retrieval for both small-molecules, complex polymeric materials, and reactions. Additionally, we show the use of chemistry foundation models in conjunction with multi-modal models such as OpenCLIP facilitate unprecedented queries and information retrieval across multiple characterization data domains. Finally, we demonstrate the integration of these models within multi-agent systems to facilitate structure and topological-based natural language queries and information retrieval for different research tasks.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: NeurIPS 2025 AI for Science Workshop.

1 Introduction

The advent of large language models (LLMs) with capabilities for tool use has enabled a dramatic expansion of LLM-based systems across chemistry and materials domains. These include applications of chatbots or multi-agent workflows to metal-organic frameworks[1, 2], inorganic materials [3], automated synthesis [4], organic synthesis [5], and protein discovery [6]. Besides the underlying agents, some of the most critical components of these workflows are the tools made available to individual agents. This can include access to retrosynthetic software [5], predictive models or computational tools [3] or generating task-specific code for automated experimentation [4] —all of which provide salient information needed by an agent to complete a given task. These examples are frequently used in combination with standard retrieval augmented generation (RAG) methodologies, where an embedded natural-language query is executed against a vector database of embedded documents to retrieve the most relevant documents and provide them to the LLM as part of its prompt [7–9]. As with other types of agent tools, this approach is effective in enhancing LLM performance for answering a particular question by providing the LLM with context it was not trained on. However, it requires significant optimization with regards to the choice of embedding model, chunk size, LLM and its available context window, question rephrasing, use of metadata filtering, and similarity metrics used for retrieval[8, 9]. The advent of large language models (LLMs) with capabilities for tool use has enabled a dramatic expansion of LLM-based systems across chemistry and materials domains. These include applications of chatbots or multi-agent workflows to metal-organic frameworks[1, 2], inorganic materials [3], automated synthesis [4], organic synthesis [5], and protein discovery [6]. Besides the underlying agents, some of the most critical components of these workflows are the tools made available to individual agents. This can include access to retrosynthetic software [5], predictive models or computational tools [3] or generating task-specific code for automated experimentation [4] —all of which provide salient information needed by an agent to complete a given task. These examples are frequently used in combination with standard retrieval augmented generation (RAG) methodologies, where an embedded natural-language query is executed against a vector database of embedded documents to retrieve the most relevant documents and provide them to the LLM as part of its prompt [8, 10, 11]. As with other types of agent tools, this approach is effective in enhancing LLM performance for answering a particular question by providing the LLM with context it was not trained on. However, it requires significant optimization with regards to the choice of embedding model, chunk size, LLM and its available context window, question rephrasing, use of metadata filtering, and similarity metrics used for retrieval [8, 11].

In many of the agent-based workflows noted above, the implemented tools may utilize structural information in the form of a SMILES string or molecular formula when performing a computation or property prediction [6, 12, 13], but not directly in RAG operations focused on structural similarity. Retrieval of relevant information based on structural similarity, whether focused small molecules, materials, or reactions, is one of the most critical tasks during any chemistry or materials research endeavor. Therefore, enabling researchers to use natural-language accompanied by chemical language to query structure-linked information resources would provide a powerful augmentation of LLM capabilities. Structure-based RAG operations necessitate a vector-based representation of a given compound or material whereby similarity queries may be conducted. While there are a significant variety of molecular fingerprints available through cheminformatics packages such as RDKit [14], the use chemistry language foundation models is highly attractive due to their potential dual use within an agentic system as an embedding model and a predictive tool. Here, the same model would be able to support both prediction tasks, such as providing accurate assessments of properties for a compound of interest and support semantic structural queries for relevant information based on a compounds structure or structural similarity. Despite these advantages, there exist only a few reports on the utilization of deep learning chemistry models as embedding models for similarity-based searches—limited largely to organic molecules and inorganic materials [15, 16]. And while the LLM systems for chemistry applications have been demonstrated to search APIs such as PubChem or the general internet [5], these search interfaces do not offer the potential breadth of both query options or structural similarity features that may be possible when using a chemistry foundation model to embed structural information and with relevant metadata. Additionally, queries based on other modalities, such as images of characterization data (e.g., NMR spectra or GPC traces), have largely not been evaluated within an LLM-based agentic system nor within traditional chemistry database interfaces in spite of the immense benefit such features would confer to researchers.

2 Results

To overcome these limitations and dramatically expand the capabilities of multi-agent systems for materials design and development tasks, we surmised that: 1) a single, high-performance chemistry foundation model could be adapted to facilitate semantic structure searches across small molecules, polymers, and reactions; 2) such a model could be used in combination with a multi-modal embedding model to enable semantic queries over images and chemical structures; and 3) these systems can be integrated within multi-agent systems to provide richer context while performing materials design tasks. Before building a complete semantic structure search-based multi-agent RAG workflow, it was imperative to evaluate the effectiveness of the latent representations generated by the chemistry language foundation models for structural similarity queries with many competent models that may fill such a role. Prior work focused on the use of ChemBERTa models to enable similarity searches for functional analogues of drugs [16–18], however there are other potential models that could work including Mol-BERT [19], Mol2Vec [20], GPT-MolBERTa [21], or ChemGPT [22]. For our investigation, we selected MoLFormer as a baseline embedding model which has recently been shown to be highly performant across numerous benchmarks from MoleculeNet [23], effective at capturing molecular similarity, as well as potentially learning 3D spatial relationships from SMILES inputs [24]. To evaluate structural similarity queries, we compiled a focused dataset of 2.5M organic small-molecules using open-source data and historical data from our own work [25–28]. The SMILES representations were canonicalized and vector embeddings were computed for each compound using MoLFormer[24] before insertion into the Milvus vector database (see Supplmentary Information for more details) [29]. Using this dataset we evaluated the ability of MoLFomer embeddings to facilitate retrieval of structurally similar compounds from sample queries on known organocatalysts for ring-opening polymerization (ROP) using the vector search capabilities inherent to Milvus (Fig. 1) [30, 31].

For the result of each query, several metrics relating to the structural similarity based on either MoLFormer embedding distance (cosine, Euclidean, Fig. 1) or fingerprints for each compound generated using RDKit (Tanimoto, RDKit, MACC, and Dice, Fig. 1). Each score represents the similarity between the resulting compound and the query compound. For 1,8-diazabicyclo[5.4.0]undec-7-ene (DBU) 1a, the top result was itself followed by several closely related structural analogues based on the cosine or Euclidean similarity (Fig. 1A, Eq. S1). Interestingly, the second rank compound **1b** was given a much lower similarity scores from most fingerprint derived metrics despite differing from **1a** only in the saturation of the imine (Fig. 1A). For **2a**, a organocatalyst for ROP designed with the aid of generative AI [31], we find that while the original query was not present in the dataset, most of the core functional group features—an cyclic 5-membered guanidine and an aromatic ring—based on visual inspection (Fig. 2B). As with **1a**, most of the fingerprint-based similarity metrics indicate the result compounds as highly dissimilar, with exception of MACC (Fig. 2B). Overall, these examples further substantiate prior results[24] demonstrating MoLFormer embeddings capture relevant structural information and similarity despite differing substantially in several cases from fingerprint-based metrics (see Supplementary Fig. S1 for an additional example query).

With the success of the similarity queries using MoLFormer embeddings in conjunction with a vector store, we sought next to expand search capabilities beyond simple small molecule queries to maximize the diversity of search options available to LLM agents. In natural language, differences in word vector embeddings have been demonstrated to capture lexical relationships between different words, with classic examples of King – Man + Woman = Queen or Paris – France + Poland = Warsaw capturing the relationships of gender and capital city[32]. Given that MoLFormer is a chemistry language model based on SMILES syntax, we surmised that differences between vector embeddings between two compounds within the base model latent space should correspond to differences in structure (Fig. 1B). Moreover, this implies that combinations of vector embeddings from compounds of interest can be used to identify novel analogues based on functional group features of both compounds as well as utilize scalar property values to manipulate the magnitude of the vector—facilitating unique similarity queries to access different sets of data. To implement these ideas, we first created an identical collection of 2.5M molecules used in Fig. 1 where instead each vector was now scaled by the compounds' corresponding molecular weight. This was tested on guanidine **3a**, which was used as a search input to either the molecular weight scaled collection, or the base collection used previously (Fig. 2A). The results in Fig. 2A demonstrated the success of both approaches in retrieving structurally relevant compounds, however, the metadata filtering

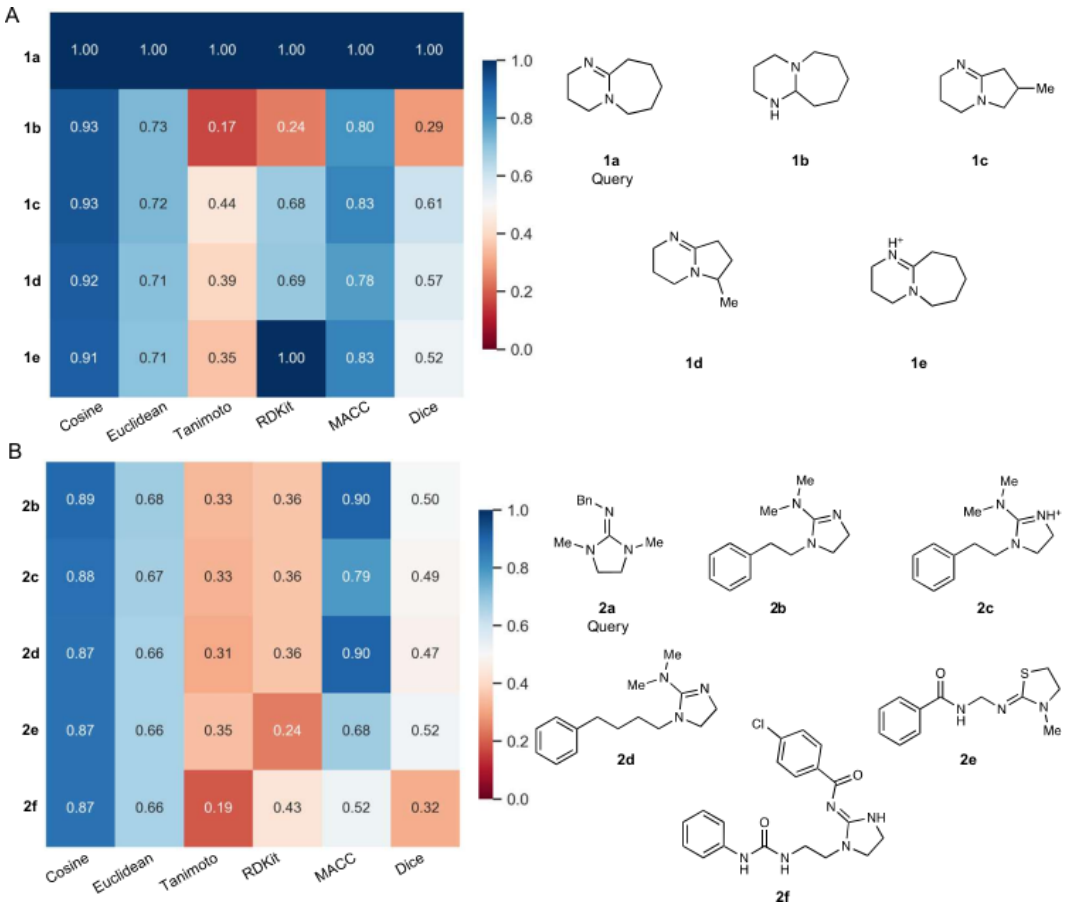


Figure 1: **A)** Heatmap of similarity metrics for the results of the query of 1a against a small-molecule collection. Structures of each of the top five closest compounds are shown beneath the heatmap. **B)** Heatmap of similarity queries of **2a** against a small-molecule collection. Structures of each of the top five closest compounds. For the similarity metrics displayed in the heatmap, cosine and Euclidean refer to the respective distance between the query embedding and the result embedding converted into similarity scores. Tanimoto, RDKit, MACC, and Dice refer to similarity scores from molecular fingerprints generated using the RDKit package for the query and returned compound. For all metrics, a value closer to 1 indicates high similarity whereas a value closer to zero indicates lower similarity. See Supplementary Information for details.

is limited to identical top results within certain molecular weight ranges (**3f**, Fig. 2A)—potentially requiring different filtering strategies depending on the desired output.

In addition to utilizing the vector magnitude to influence similarity search results, we also investigated whether the addition and subtraction of vectors corresponding to different functional groups resulted in similar lexical relationships as observed with examples like: King – Man + Woman = Queen. This approach was evaluated on two catalysts **4a** and **4b** where their corresponding vector embeddings were subjected such operations. With **4a** the subtraction of the vector Corresponding to dimethylurea and the addition of dimethylthiourea resulted in a vector embedding that provided corresponding thioureas (**4c**) when queried against the database (Fig. 2B). With **4b**, the results of similar operation were less obviously successful, but this is anticipated behavior based on a limited collection of molecules (Fig. 2B). Nonetheless, the results did provide similar compounds with fluorine containing amines. Finally, the vector embeddings of **4a** and **4b** were averaged and queried against the database with the top results bearing the structural features of both the parent molecules (Fig. 2C).

Having demonstrated the effectiveness of using MoLFormer vector embeddings on a variety of different similarity search strategies, we next sought to evaluate their feasibility towards polymers and reactions. MoLFormer was trained compounds with SMILES strings containing less than 200 tokens

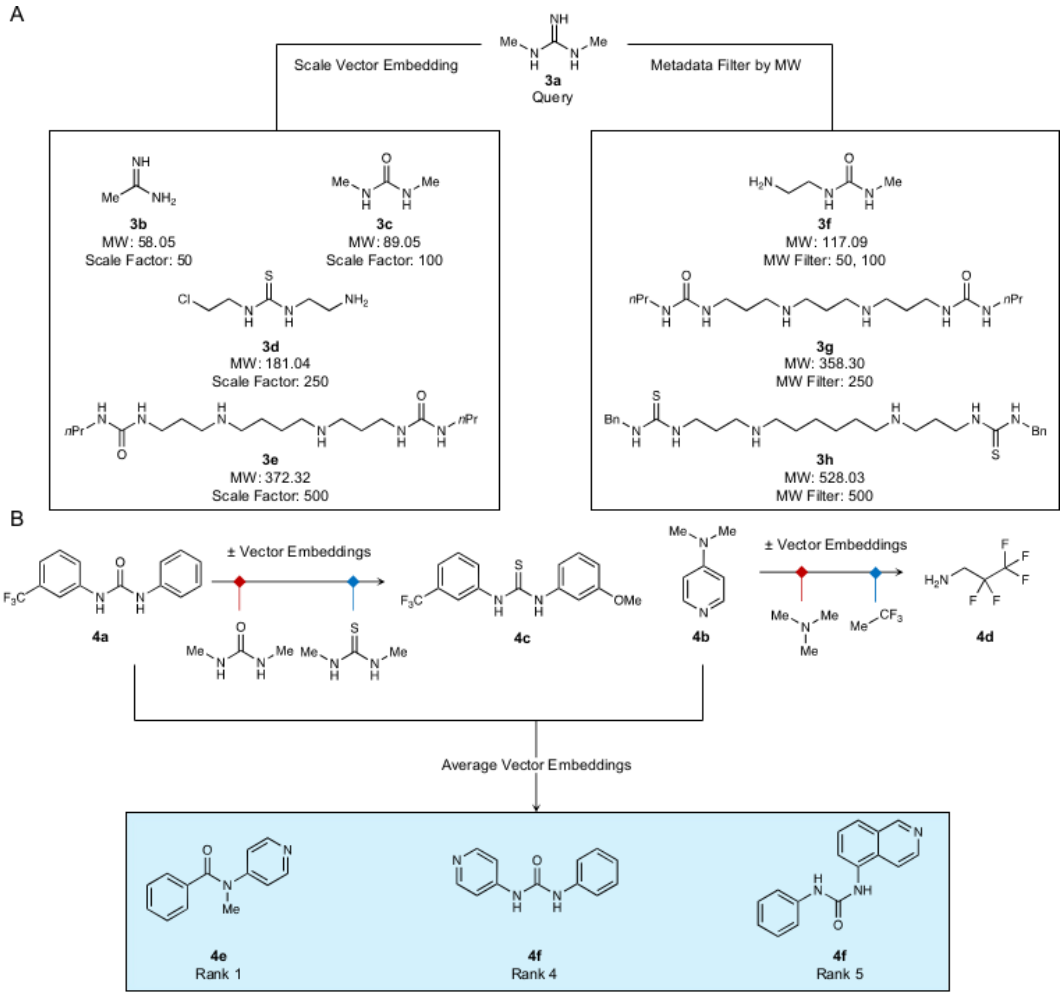


Figure 2: **A)** Comparison of results obtained for queries using run using metadata filters compared to normalized vector embeddings scaled by a target molecular weight. **B)** Schematic depiction of search results obtained through mathematical manipulation of query vector embeddings through addition-subtraction of corresponding functional group vectors or averaging two queries.

were filtered from the pre-training dataset, precluding large SMILES strings which may potentially represent macromolecules[24]. Outside of biological materials, such as peptides, nucleic acids or large natural products, macromolecules and other complex materials often exhibit stochastic features relating to their structure and are hence poorly represented by discrete, one-dimensional SMILES strings. This problem is frequently circumvented in deep learning models through a reduction in the complexity by treating polymers as single, discrete repeating units via SMILES strings with variable attachment points denoted by the asterisk character (PSMILES) [33, 34]. While this approach tends to overlook the stochastic nature of materials as well as neglects to account for end groups or more complex polymer topologies, it does produce systems capable of providing predicted values for polymer properties within these restrictions [33, 34]. As the tokenizer for MoLFormer covers the entire SMILES grammar [24, 35], including the use of special tokens such as the use of asterisks for variable attachment points, we hypothesized that it could also serve as a suitable source of embeddings to facilitate queries based on both polymer structural and topological similarity. It was unclear, however, as to whether these latent embeddings would capture the same level of molecular similarity given SMILES fragments with asterisks were unlikely to have been part of any pre-training set for the model[24]. With this in mind, we leveraged dot-separated SMILES strings containing asterisks to represent repeat units or other polymer components to evaluate their suitability for use in semantic queries for polymeric materials.

To evaluate the use of MoLFormer embeddings applied to polymers, we first curated a set of polymer data from both open literature and our own historical data totaling in 2.5M PSMILES strings representing predominantly homopolymers [36–41]. As with the small-molecule embeddings, these were inserted into a Milvus collection, against which similarity queries may be run. First, we tested the embeddings on very simple queries such as polyimide **S4a** and polystyrene **S4c**, both of which returned sets of highly similar compounds (Fig. S3A). Averaging the embeddings of **S4a** and **S4c** enabled the search to identify features of both, consistent with results observed on the small-molecules and again indicative of the ability of MoLFormer embeddings to capture semantic structural relationships. Additional examples of basic polymer similarity searches using PSMILES can be found in Fig. S2.

Next, we investigated the possibility of utilizing additional scalar weights with the component vector embeddings to enable not only querying for structural similarity but also topological and molecular weight similarity. In this approach, the polymer structure is broken into its structural components, each of which is embedded independently using MoLFormer (Fig. 3A). The individual embeddings are then subjected to weighted averaging method to afford the final vector embedding for a polymer (Fig. 3A). To evaluate this approach, we generated a synthetic dataset consisting of both ROP and acrylate polymers (2M) with randomly generated molecular weights, dispersity, and degrees of polymerization. Two approaches for computing the final polymer embedding from the component embeddings were investigated. The first approach averaged the individual component vectors and then scaled the averaged embedding by the dispersity and the $\log_{10}$ of the M_n (see Supplementary Information). The second approach focused on the log weighted average of each of the structural components' contribution to the overall M_n of the polymer. Two separate collections from synthetic polymer data were constructed utilizing these embedding approaches and the same query polymer (**5a**, Fig. 3A) was executed against both collections using their respective embedding methods. For the first embedding approach (Fig. 3B), the returned results exhibited high cosine similarity yet diverged strongly in terms of Euclidean similarity measures (Fig. 3B). This can readily be seen within returned results with **5b** being identical structurally but differing more starkly in terms of molecular weight (Fig. 3B). The next closest compounds **5c-1** and **5c-2** exhibited some similarity in structure and in terms of molecular weight. In contrast, the second embedding approach provided results with high cosine and Euclidean similarity (Fig. 3B). The first three results (**5f**) were both structurally identical to the query—with the top result (5f-1) being closest in M n —and the only examples of 5a in the entire collection (Fig. 3C). Additionally, following results both contained the same structural components within the polymer (**5g** and **5h**, Fig. 3C), as opposed to the other embedding methodology, where only the methanol initiator component was conserved across all the top results (Fig. 3D). Overall, these results demonstrate both the strong influence of the methodology used to scale embeddings on the returned results as well as the ability to use embeddings in combination with a scalar of property data to facilitate unprecedented similarity queries for polymeric materials.

As with polymers, reactions—represented using reaction SMILES syntax[42]—has not been examined directly using MoLFormer and despite the existence of many transformer-based models for reactions [35, 43], we were interested in probing the versatility of MoLFormer embedding model reaction similarity queries. Based on our results with polymeric systems, we anticipated reactions to behave in a similar manner given a similar syntactical construction. The data for the evaluation was sourced from an open 2M reaction dataset sourced from the USPTO used previously in transformer models for reactions, open publications, and historical experimental polymerization reaction data [44]. In addition to querying on whole reactions, we also focused on queries involving one to two reagents (Fig. S4A). In these examples, **S5a**, **S5b**, and their average vector where each used to search the database and provided reaction examples where structurally related reagents were used or the query molecules themselves (Fig. S4A). Finally, testing the order of **S5a** and **S5b**, either in a dot separated substrate series or on either side of the double angle brackets (») separating substrates and products produced markedly different results, indicating the importance of order within the reaction SMILES sequence. The difference in results when angle brackets are used is understandable considering that this would indicate two very different reactions when the order is reversed. However, unlike with polymers, the distinction between the ordering of reactants or products in reaction SMILES has less relevance unless some order of addition is being encoded. In total, these results indicate that the base model of MoLFormer is well-suited to capture structural relationships among small-molecules, polymers, and reactions.

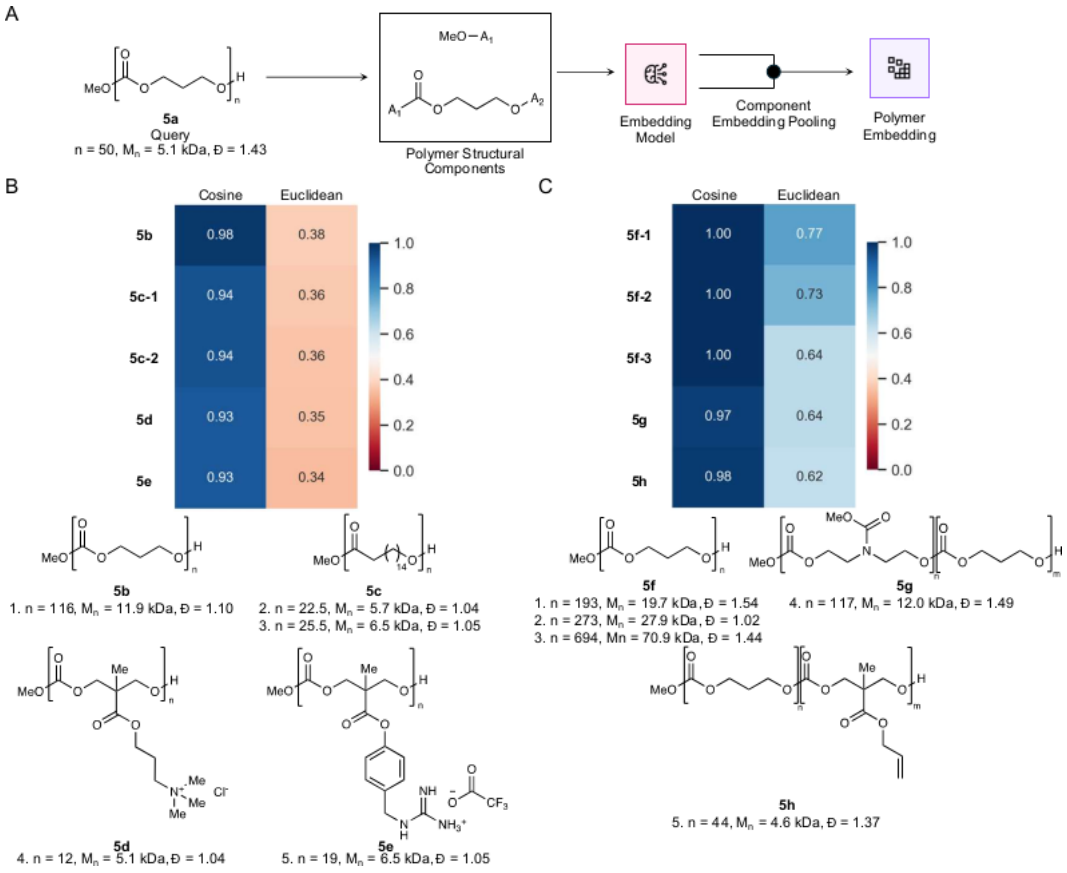


Figure 3: **A)** Schema of how an example query polymer **5a** is broken into structural components and then processed into single embedding to facilitate querying across both polymer structure and other characteristics such as M_n. **B)**. Results from query of **5a** embedded using equation S2. **C)** Results from query of **5b** embedded using equation S3.

The demonstration of MoLFormer as an effective embedding model beyond simple small molecules prompted us to examine the use of chemical structural embeddings with additional modalities. Association of chemical structure with different types of characterization data is a highly important task for any project within chemistry and materials domains. While characterization data is available in many formats, we opted to leverage the data pre-plotted and saved as images as input for creation of their corresponding vector embeddings. The ability to query available characterization data by either image or structure would be a powerful addition to both traditional data infrastructures and RAG pipelines. Implementing this task would typically require either some form contrastive learning or latent fusion, necessitating fine-tuning of an existing model or training of an additional model. Instead, we opted to test alternative strategies where the chemical components of a particular piece of characterization data would be embedded using MolFormer while the image components would be embedded using OpenCLIP (see Supplementary Information for details) [45, 46]. While CLIP models for images can also co-embed text captions, it is not anticipated these embeddings would be able to effectively capture nuances in chemical structure as a chemistry focused model like MoLFormer. Instead, the text embedder of the OpenCLIP model can be used to embed information captured as a natural language caption, which can be automatically generated from the metadata of characterization data files, adding additional natural language query capabilities. To evaluate this approach, we compiled a small dataset of labeled images of ^{1}H, ^{13}C, and ^{19}F NMR spectra (see Supplementary Information). The labels for the spectra included their corresponding chemical components as SMILES strings and natural language captions generated from the spectra metadata. From this dataset we can test both image-based and structure-based queries on the embedded dataset with excellent results across numerous collection organization strategies (Fig. 4). Despite there being only minor differences between the image features plots—colors, whitespace, axes, and image size—of the

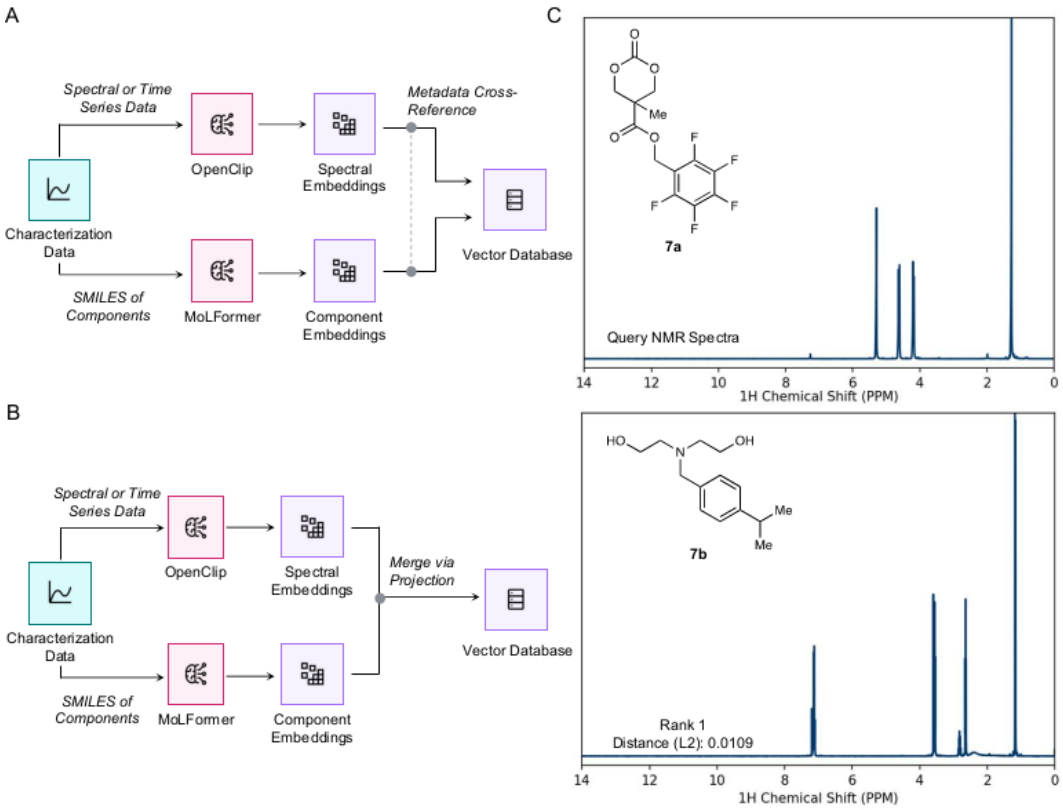


Figure 4: **A)** Schematic of embedding process for characterization data where the image and structural components share a metadata reference. **B)** Schematic of embedding process for characterization data where the image and structural embeddings are projected into a single vector before storage in a vector database. C) Example of query ^{1}H NMR spectra and the top rank result from a querying against vector store built using cross-referenced metadata approach.

embedded plots, OpenCLIP can effectively differentiate between different characterization techniques without additional metadata filtering on the query. This implies that the differentiation is likely occurring due to shape of the plotted traces or spectra (Fig. 4).

The use of chemistry LLMs such as MoLFormer and multi-modal models like OpenCLIP, in combination with post-embedding pooling and compression strategies, has facilitated the creation of a variety of vector stores which in turn can support a large variety of similarity queries. This alone would make a powerful addition to traditional database architectures and search capabilities, yet it's the connection of such vector databases within larger agentic RAG workflows can offer the potential of significant time savings in complex research tasks requiring the merging and summarization of data retrieved from complex structure, image, and natural language-based queries. In this regard, access to different vector stores and their respective embedding strategies is provided to LLM agents as tools which may be used in the context of a particular task. We utilized the LangGraph library (v 0.1.19) [47] to develop a hierarchical multi-agent adaptive self-reflective RAG system (Fig. 5A and 5B), which takes a question as input and outputs, the solution as a formatted report. As shown in Fig. 5, the hierarchical workflow is directed by a supervisor agent which leverages query analysis [48] to adaptively route a user's question to the correct worker agent. The system consists of four worker agents, which specialize in small molecules, polymers, chemical reactions, or NMR spectra. Each of these workers autonomously implement a multi-agent-based self-reflective RAG workflow [49]. As detailed in Fig. 5B, the workflow consists of several steps including retrieving and evaluating documents, generating responses using retrieved documents, revising user input to improve retrieved documents, checking generated responses for hallucinations and verifying each response completely addresses a user's question. Each step of this process is autonomously directed by the worker agents, which runs until all checks are passed. Once all checks are passed, the answer is sent to the report

generation tool, which summarizes the agent's findings (Fig. S5 and S6). This report also summarizes the content retrieved from the vector store used by the agent to generate the report in order to improve the overall transparency of the workflow.

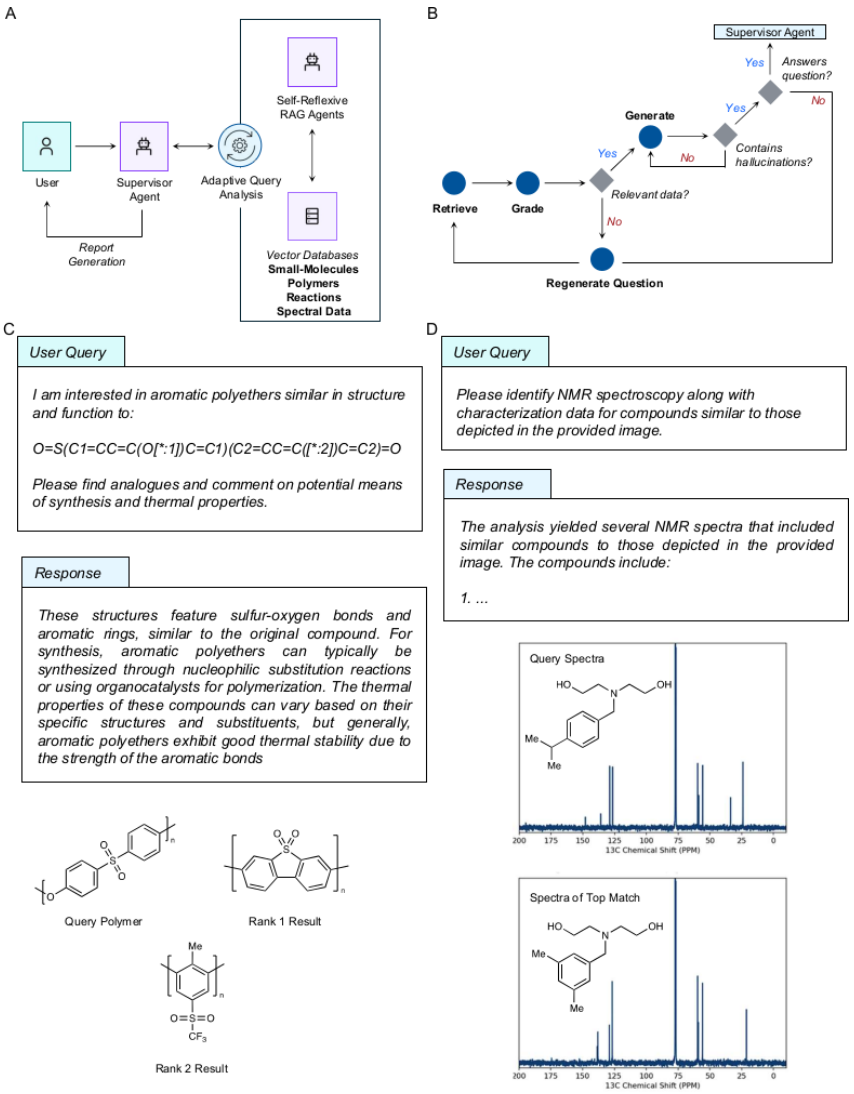


Figure 5: **A)** Schematic overview of hierarchical multi-agent system RAG to handle queries for different collections. **B)** Self-reflective RAG pipeline used in the adaptive query analysis by each RAG worker agent. **C)** Summarized example of a polymer query and response from the multi-agent system. **D)** Summarized example of an NMR spectrum query and response from the multi-agent.

This workflow was demonstrated on a variety of tasks, each of which was able to effectively utilize the structure and/or image-based vector store tools in combination with other natural language-based vector databases from historical manuscripts. Summarized examples are shown in Fig. 5C and 5D with the full reports shown in Supplementary Figs. S5 and S6. As shown in Fig. 5C, the multi-agent workflow was able to successfully complete the user query using self-reflective RAG for polymers returning two similar analogs and provided additional content to justify the selection of these analogs. It should be noted that the collections utilized for the evaluation with the multi-agent system were limited to 250k examples from the original collection, thus covering a narrower range of compound types. Regardless, the structure-based RAG approach was able to retrieve close analogues for further utilization by the agents. For Fig. 5D, the agentic system was asked to identify similar ^{13}C NMR spectra to the one of a diethanolamine-based carbonate monomer precursor [50]. The question was routed to the multi-modal NMR self-RAG agent who reviewed and summarized the characterization

data of four ^{13}C NMR spectra images (three retrieved images and the image corresponding to the input) summarizing the location of the peaks in these images that are similar to the input NMR spectra. The report also provides a visualization of the input image, which enables easier review and interpretation of the provided results. Both reports were reviewed by a domain expert who confirmed the validity of the findings. In this example, the agent was not instructed to filter the RAG results if they contained the original query image, and was still able to successfully find several spectra of closely related diethanolamine-based cyclic carbonates or their precursors (Fig. 5C, Fig. S6).

As a final demonstration of the multi-modal agentic workflow, we implemented a user interface (UI) connection using Open WebUI, an open-source UI for chat-based interactions[51]. Additionally, we connected to the workflow foundation models for chemistry and materials property prediction[52, 53], which provided additional context for each retrieved document in the self-reflexive RAG pipeline. Figs. S7-S10 show examples of a polymer- and characterization data-focused queries, showing screenshots from the workflow responses and context summaries. The multi-modal characterization data queries were expanded to include HSQC, EI-MS, and IR data using a open-source synthetic dataset[54].

3 Conclusion

The accelerated development of novel materials and catalysts necessitates dramatic improvement of human–AI interactions to facilitate both effective co-designs as well as realistic implementations within an experimental setting. Multi-agent RAG systems integrated with chat interfaces hold significant promise to become useful assistants for researchers operating in laboratory settings. Here, we have demonstrated that chemical foundation models coupled with powerful Multi-modal models can facilitate unique types of structural or architectural focused queries on small-molecules, polymers, and reactions. This represents a significant enhancement of search capabilities not typically found in traditional database systems for chemistry research. Moreover, the coupling of these systems within LLM-based multi-agent system can provide a significant advantage for reducing the time needed to retrieve and summarize relevant multi-modal structure-based information commonly required across all material research projects.

4 Code and Data Availability

Data and code will be made available upon final publication.

5 Conflict of Interest

A patent application on aspects of multimodal semantic search for materials has been filed by IBM (application number 18/798970, inventors: N.H.P, T.E., and J.L.H). N.H.P, J.L.H, T.E, and S.C are employees of IBM.

6 Acknowledgments

This material is based upon work supported by the NSF under Grant No. DBI-1548297.

References

[1] Y. Kang and J. Kim, "Chatmof: an artificial intelligence system for predicting and generating metal-organic frameworks using large language models," *Nature communications*, vol. 15, no. 1, p. 4705, 2024.

[2] Z. Zheng, O. Zhang, C. Borgs, J. T. Chayes, and O. M. Yaghi, "Chatgpt chemistry assistant for text mining and the prediction of mof synthesis," *Journal of the American Chemical Society*, vol. 145, no. 32, pp. 18048–18062, 2023.

[3] Y. Chiang, E. Hsieh, C.-H. Chou, and J. Riebesell, "LLaMP: Large Language Model Made Powerful for High-fidelity Materials Knowledge Retrieval and Distillation," 2024. arXiv:2401.17244 [cond-mat].

[4] D. A. Boiko, R. MacKnight, B. Kline, and G. Gomes, "Autonomous chemical research with large language models," *Nature*, vol. 624, no. 7992, pp. 570–578, 2023.

[5] A. M. Bran, S. Cox, O. Schilter, C. Baldassari, A. D. White, and P. Schwaller, "Augmenting large language models with chemistry tools," *Nature Machine Intelligence*, vol. 6, no. 5, pp. 525–535, 2024.

[6] A. Ghafarollahi and M. J. Buehler, "ProtAgents: protein discovery via large language model multi-agent collaborations combining physics and machine learning," *Digital Discovery*, vol. 3, no. 7, pp. 1389–1409, 2024.

[7] H. Li, Y. Su, D. Cai, Y. Wang, and L. Liu, "A survey on retrieval-augmented text generation," 2022. arXiv:2202.01110 [cs].

[8] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in neural information processing systems*, vol. 33, pp. 9459–9474, 2020.

[9] P. Zhao, H. Zhang, Q. Yu, Z. Wang, Y. Geng, F. Fu, L. Yang, W. Zhang, J. Jiang, and B. Cui, "Retrieval-Augmented Generation for AI-Generated Content: A Survey," 2024. arXiv:2402.19473 [cs].

[10] C. Liao, Y. Yu, Y. Mei, and Y. Wei, "From Words to Molecules: A Survey of Large Language Models in Chemistry," 2024. arXiv:2402.01439 [cs, q-bio].

[11] P. Zhao, H. Zhang, Q. Yu, Z. Wang, Y. Geng, F. Fu, L. Yang, W. Zhang, J. Jiang, and B. Cui, "Retrieval-augmented generation for ai-generated content: A survey," *arXiv preprint arXiv:2402.19473*, 2024.

[12] M. J. Buehler, "Generative Retrieval-Augmented Ontologic Graph and Multiagent Strategies for Interpretive Large Language Model-Based Materials Design," *ACS Engineering Au*, vol. 4, no. 2, pp. 241–277, 2024.

[13] A. Ghafarollahi and M. J. Buehler, "AtomAgents: Alloy design and discovery through physics-aware multi-modal multi-agent artificial intelligence," 2024. arXiv:2407.10022 [cond-mat].

[14] D. Boldini, D. Ballabio, V. Consonni, R. Todeschini, F. Grisoni, and S. A. Sieber, "Effectiveness of molecular fingerprints for exploring the chemical space of natural products," *Journal of Cheminformatics*, vol. 16, no. 1, p. 35, 2024.

[15] Y. Yang, M. Liu, and J. R. Kitchin, "Neural network embeddings based similarity search method for atomistic systems," *Digital Discovery*, vol. 1, no. 5, pp. 636–644, 2022.

[16] C. W. Kosonocky, A. L. Feller, C. O. Wilke, and A. D. Ellington, "Using alternative SMILES representations to identify novel functional analogues in chemical similarity vector searches," *Patterns*, vol. 4, no. 12, p. 100865, 2023.

[17] S. Chithrananda, G. Grand, and B. Ramsundar, "ChemBERTa: Large-scale self-supervised pretraining for molecular property prediction," 2020. arXiv:2010.09885 [physics, q-bio].

[18] W. Ahmad, E. Simon, S. Chithrananda, G. Grand, and B. Ramsundar, "ChemBERTa-2: Towards Chemical Foundation Models," 2022. arXiv:2209.01712 [cs, q-bio].

[19] J. Li and X. Jiang, "Mol-BERT: An effective molecular representation with bert for molecular property prediction," *Wireless Communications and Mobile Computing*, vol. 2021, no. 1, p. 7181815, 2021.

[20] S. Jaeger, S. Fulle, and S. Turk, "Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition," *Journal of Chemical Information and Modeling*, vol. 58, no. 1, pp. 27–35, 2018.

[21] S. Balaji, R. Magar, Y. Jadhav, and A. B. Farimani, "GPT-MolBERTa: GPT Molecular Features Language Model for molecular property prediction," 2023. arXiv:2310.03030 [physics].

[22] N. C. Frey, R. Soklaski, S. Axelrod, S. Samsi, R. Gómez-Bombarelli, C. W. Coley, and V. Gadepally, "Neural scaling of deep chemical models," *Nature Machine Intelligence*, vol. 5, no. 11, pp. 1297–1305, 2023.

[23] Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, and V. Pande, "MoleculeNet: a benchmark for molecular machine learning," *Chemical Science*, vol. 9, no. 2, pp. 513–530, 2018.

[24] J. Ross, B. Belgodere, V. Chenthamarakshan, I. Padhi, Y. Mroueh, and P. Das, "Large-scale chemical language representations capture molecular structure and properties," *Nature Machine Intelligence*, vol. 4, no. 12, pp. 1256–1264, 2022.

[25] S. Gallarati, P. van Gerwen, R. Laplaza, S. Vela, A. Fabrizio, and C. Corminboeuf, "Oscar: an extensive repository of chemically and functionally diverse organocatalysts," *Chemical Science*, vol. 13, no. 46, pp. 13782–13794, 2022.

[26] B. C. Haas, M. A. Hardy, S. S. SV, K. Adams, C. W. Coley, R. S. Paton, and M. S. Sigman, "Rapid prediction of conformationally-dependent dft-level descriptors using graph neural networks for carboxylic acids and alkyl amines," *Digital Discovery*, vol. 4, no. 1, pp. 222–233, 2025.

[27] J. Zheng, "IUPAC/Dissociation-Constants: v1.0." `https://zenodo.org/records/7236453`. Accessed: 2024-08-12.

[28] J. Wu, Y. Wan, Z. Wu, S. Zhang, D. Cao, C.-Y. Hsieh, and T. Hou, "MF-SuP-pKa: Multi-fidelity modeling with subgraph pooling mechanism for pka prediction," *Acta Pharmaceutica Sinica B*, vol. 13, no. 6, p. 2572–2584, 2023.

[29] J. Wang, X. Yi, R. Guo, H. Jin, P. Xu, S. Li, X. Wang, X. Guo, C. Li, X. Xu, K. Yu, Y. Yuan, Y. Zou, J. Long, Y. Cai, Z. Li, Z. Zhang, Y. Mo, J. Gu, R. Jiang, Y. Wei, and C. Xie, "Milvus: A Purpose-Built Vector Data Management System," in *Proceedings of the 2021 International Conference on Management of Data*, SIGMOD '21, (New York, NY, USA), pp. 2614–2627, Association for Computing Machinery, 2021.

[30] M. K. Kiesewetter, E. J. Shin, J. L. Hedrick, and R. M. Waymouth, "Organocatalysis: opportunities and challenges for polymer synthesis," *Macromolecules*, vol. 43, no. 5, pp. 2093–2107, 2010.

[31] N. H. Park, M. Manica, J. Born, J. L. Hedrick, T. Erdmann, D. Y. Zubarev, N. Adell-Mill, and P. L. Arrechea, "Artificial intelligence driven design of catalysts and materials for ring opening polymerization using a domain-specific language," *Nature Communications*, vol. 14, no. 1, p. 3686, 2023.

[32] E. Vylomova, L. Rimell, T. Cohn, and T. Baldwin, "Take and Took, Gaggle and Goose, Book and Read: Evaluating the Utility of Vector Differences for Lexical Relation Learning," 2016. arXiv:1509.01692 [cs].

[33] C. Xu, Y. Wang, and A. Barati Farimani, "Transpolymer: a transformer-based language model for polymer property predictions," *npj Computational Materials*, vol. 9, no. 1, p. 64, 2023.

[34] C. Kuenneth and R. Ramprasad, "polyBERT: a chemical language model to enable fully machine-driven ultrafast polymer informatics," *Nature Communications*, vol. 14, no. 1, p. 4099, 2023.

[35] P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas, and A. A. Lee, "Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction," *ACS Central Science*, vol. 5, no. 9, pp. 1572–1583, 2019.

[36] R. Ma and T. Luo, "PI1M: a benchmark database for polymer informatics," *Journal of Chemical Information and Modeling*, vol. 60, no. 10, pp. 4684–4690, 2020.

[37] M. Aldeghi and C. W. Coley, "A graph representation of molecular ensembles for polymer property prediction," *Chemical Science*, vol. 13, no. 35, pp. 10486–10498, 2022.

[38] I. V. Volgin, P. A. Batyr, A. V. Matseevich, A. Y. Dobrovskiy, M. V. Andreeva, V. M. Nazarychev, S. V. Larin, M. Y. Goikhman, Y. V. Vizilter, A. A. Askadskii, *et al.*, "Machine learning with enormous "synthetic" data sets: predicting glass transition temperature of polyimides using graph convolutional neural networks," *ACS Omega*, vol. 7, no. 48, pp. 43678–43691, 2022.

[39] J. Hu, Z. Li, J. Lin, and L. Zhang, "Prediction and interpretability of glass transition temperature of homopolymers by data-augmented graph convolutional neural networks," *ACS Applied Materials & Interfaces*, vol. 15, no. 46, pp. 54006–54017, 2023.

[40] J. Yang, L. Tao, J. He, J. R. McCutcheon, and Y. Li, "Machine learning enables interpretable discovery of innovative polymers for gas separation membranes," *Science Advances*, vol. 8, no. 29, p. eabn9545, 2022.

[41] M. Ohno, Y. Hayashi, Q. Zhang, Y. Kaneko, and R. Yoshida, "Smipoly: generation of a synthesizable polymer virtual library using rule-based polymerization reactions," *Journal of Chemical Information and Modeling*, vol. 63, no. 17, pp. 5539–5548, 2023.

[42] D. Weininger, "SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules," *Journal of Chemical Information and Computer Sciences*, vol. 28, p. 31–36, Feb. 1988.

[43] A. M. Bran and P. Schwaller, "Transformers and large language models for chemistry and drug discovery," in *Drug Development Supported by Informatics*, pp. 143–163, Springer, 2024.

[44] P. Schwaller, D. Probst, A. C. Vaucher, V. H. Nair, D. Kreutter, T. Laino, and J.-L. Reymond, "Mapping the space of chemical reactions using attention-based neural networks," *Nature Machine Intelligence*, vol. 3, no. 2, pp. 144–152, 2021.

[45] G. Ilharco, M. Wortsman, N. Carlini, R. Taori, A. Dave, V. Shankar, H. Namkoong, J. Miller, H. Hajishirzi, A. Farhadi, and L. Schmidt, "OpenCLIP." `https://zenodo.org/records/5143773`, 2021. Accessed: 2024-08-12.

[46] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," in *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, PMLR, 2021.

[47] "LangGraph." `https://github.com/langchain-ai/langgraph`. Accessed: 2024-08-21.

[48] S. Jeong, J. Baek, S. Cho, S. J. Hwang, and J. C. Park, "Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity," 2024. arXiv:2403.14403 [cs].

[49] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, "Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection," 2023. arXiv:2310.11511 [cs].

[50] J. L. Hedrick, V. Piunova, N. H. Park, T. Erdmann, and P. L. Arrechea, "Simple and efficient synthesis of functionalized cyclic carbonate monomers using carbon dioxide," *ACS Macro Letters*, vol. 11, no. 3, pp. 368–375, 2022.

[51] "Open WebUI." `https://github.com/open-webui/open-webui`, Aug. 2025. Accessed: 2025-03-01.

[52] E. Soares, N. Park, E. V. Brazil, and V. Y. Shirasuna, "A large encoder-decoder polymer-based foundation model," in *AI for Accelerated Materials Design - NeurIPS 2024*, 2024.

[53] E. Soares, V. Shirasuna, E. V. Brazil, R. Cerqueira, D. Zubarev, and K. Schmidt, "A large encoder-decoder family of foundation models for chemical language," 2024. arXiv:2407.20267 [physics].

[54] M. Alberts, O. Schilter, F. Zipoli, N. Hartrampf, and T. Laino, "Unraveling molecular structure: A multimodal spectroscopic dataset for chemistry," *Advances in Neural Information Processing Systems*, vol. 37, p. 125780–125808, 2024.

[55] "HuggingFace–ibm/MoLFormer-XL-both-10pct." `https://huggingface.co/ibm/MoLFormer-XL-both-10pct`. Accessed: 2024-08-12.

A Supplementary Material

Vector Database Setup. All SMILES embeddings were computed using the *ibm/MoLFormer-XL-both-10pct* model available through HuggingFace using canonicalized SMILES (with the exception of the USPTO dataset, which was used as is) [55]. For the example queries in Figures 1-4, the vector embeddings for each compound were L2 normalized prior to insertion into a Milvus lite vector database using and Hierarchical navigable small world (HNSW) index and the L2 distance as the metric. For the multi-modal examples (Fig. 5) SMILES embeddings were not L2 normalized and used in a Milvus standalone vector database using *IVF_FLAT* index and the L2 distance metric. For spectra containing more than one identifiable compound, including the NMR solvent, their corresponding embeddings were averaged prior to insertion. Image embeddings were computed using OpenCLIP via the LangChain library. The model used was *ViT-g-14* with the *laion2b_s34b_b88k* checkpoint. Image embeddings were stored in a separate Milvus collection, using an *IVF_FLAT* index and L2 distance metric, and cross-referenced with their corresponding compound embeddings.

Similarity Metrics. For small molecule and polymer example queries, similarity metrics were computed on the basis of both the MoLFormer embeddings and molecular fingerprints computed using RDKit. Cosine similarity was computed by subtracting the cosine distance between the query and result embeddings (measured using the SciPy library) from Eq.1. Euclidean similarity (E_s) was computed by the following equation:

$$E_s = 1/(1 + E_d) \tag{1}$$

where E_d is the Euclidean distance between the query and result embeddings. Tanimoto and Dice similarities were computed using the corresponding query and result Morgan fingerprints with a radius of 2 and a dimension of 2048. RDKit similarity was computed using the built-in RDKit fingerprints of the query and compound. Molecular ACCess System (MACCS) similarity was computed using the MACCS Keys fingerprints of the query and the compound using the RDKit package.

Language Agent Network. As noted in the main text, the multi-agent framework was assembled using LangChain and LangGraph libraries. The supervisor agent utilized *GPT-4o* mini model while other agents used either *llava 7b* or *llama3.1 8b* models. All Milvus collections were instantiated as separate vector stores with customized embedding functions with either MoLFormer or OpenCLIP, prior to connection with LLM agents as retrievers. Full code for agent network will be released in both subsequent drafts of preprint and final publication.

Synthetic Polymer Data Generation. A synthetic dataset of homopolymers and AB block copolymers of polycarbonates, polyesters, polysiloxanes, polyacrylates, and polystyrenes was created through enumerative combination of compatible monomers and initiators pulled from historical datasets. Values for M_n, DP_n, and dispersity where randomly generated within specific ranges and assigned to the enumerated polymer structure. The resulting dataset contained >1M synthetic polymers with attendant molecular weight properties.

Vector embeddings for the synthetic polymer data were computed using the following equations:

$$\vec{x} = \log_{10} M_n \cdot D \cdot \frac{1}{n} \sum_{i=1}^{n} \vec{v}_i \tag{2}$$

$$\vec{x} = \frac{1}{n} \sum_{i=1}^{n} \log_{10} x_i \cdot \vec{v}_i \tag{3}$$

where $\vec{x}$ is the final vector embedding, $\vec{v}$ is the MolFormer embedding of the individual polymer component, M_n is the number average molecular weight, D is the dispersity of the polymer, and x_i is the weight fraction contribution of the component to the total M_n.

Open WebUI Interface. A variation of the agentic workflow shown in Fig. 5 implemented in LangGraph was connected to an Open WebUI interface via the *pipelines* API[51], allowing execution of the custom, multi-modal RAG pipeline and streaming of custom events emitted from the workflow. Predictive models for polymer properties used fine-tuned checkpoints from a currently unreleased foundation model for polymers[52], with representation interconversion handled during a query processing step. In this agentic workflow, *granite3.2:8b* and *llama3.1* models were used for retrieval, analysis, summarization and routing tasks.

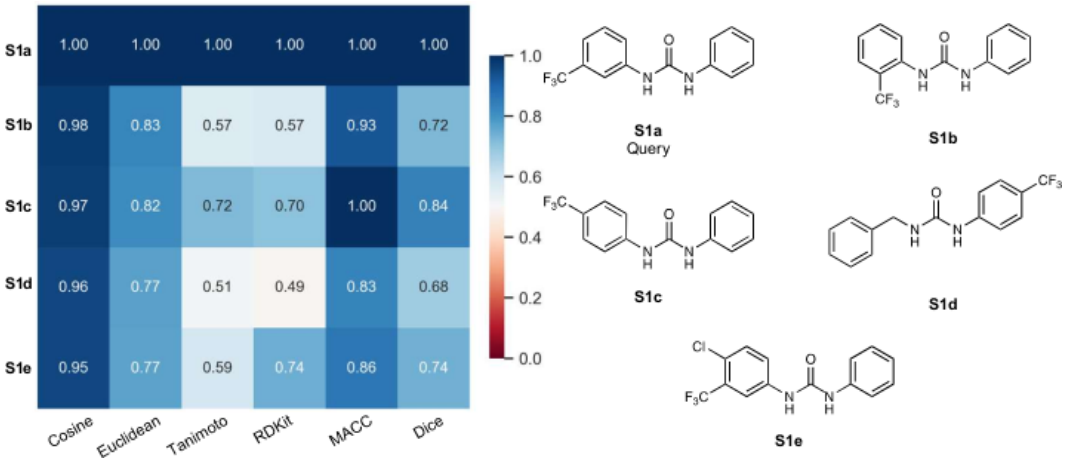


Fig. S1: Additional example of small-molecule similarity query.

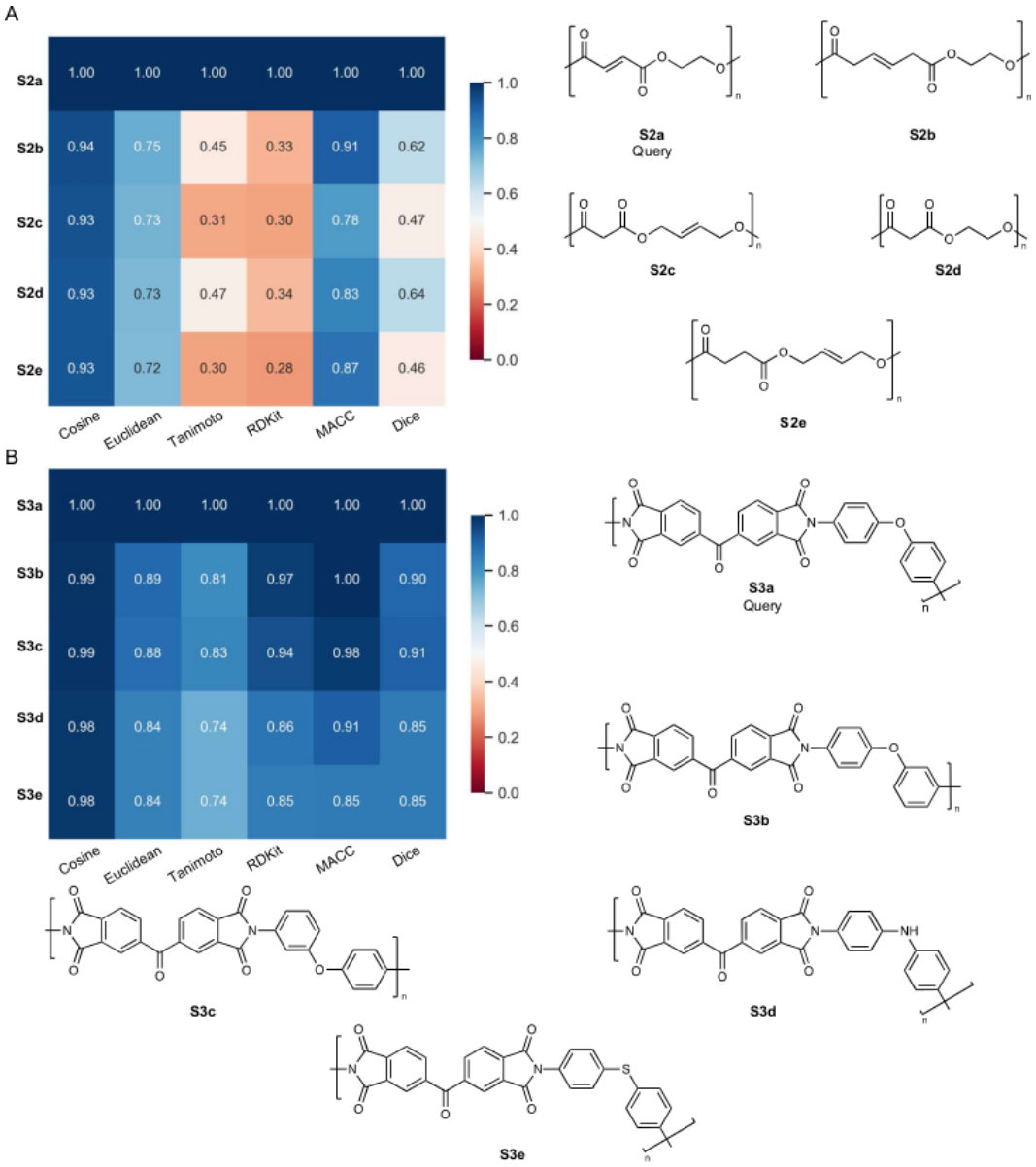


Fig. S2: Examples of similarity searches for polymers based on PSMILES strings.

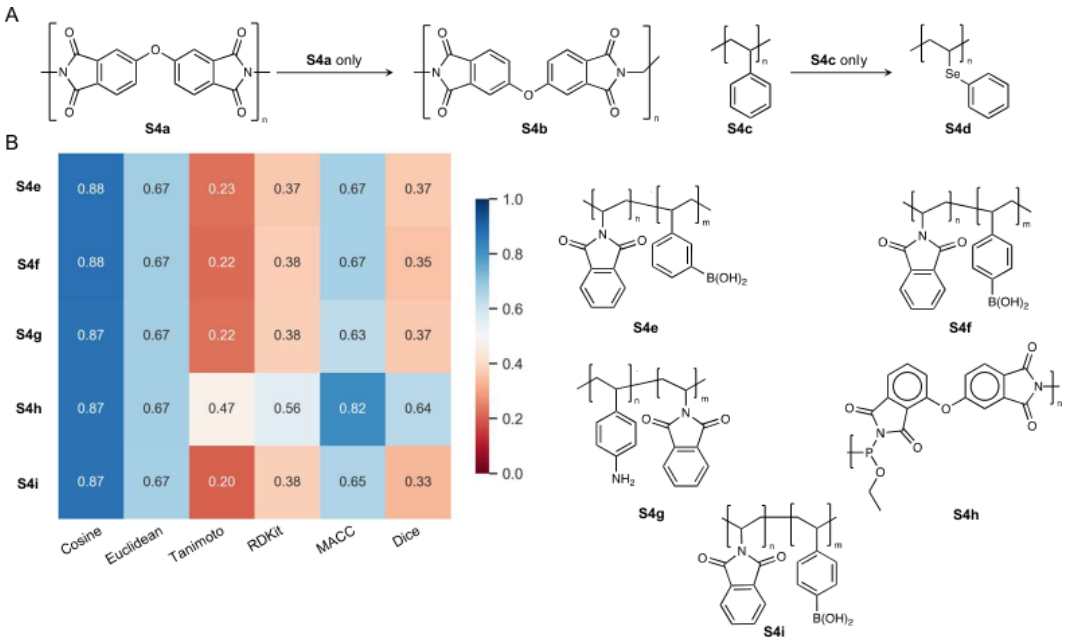


Fig. S3: **A)** Examples of similarity searches for polymers based on PSMILES strings. **B)** Results of a query on the average embedding of **S4a** and **S4b**

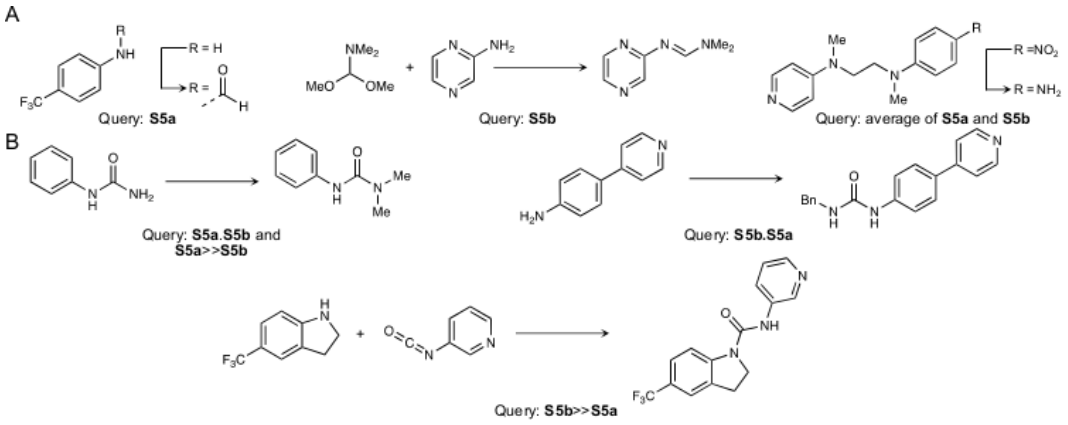
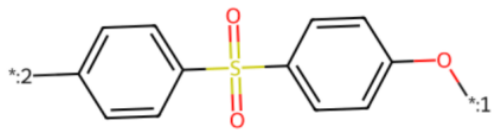


Fig. S4: **A)** Top results of queries against 2.5 M USPTO reaction collection of embedded reaction SMILES for either **S5a**, **S5b**, or their average. **B)** Comparison of top results from concatenated query SMILES strings of S5a andS5b either dot separated, such as in **S5a**. **S5b**, or angle-bracket separated, such as in **S5a**»**S5b**. Other reagents and solvents are excluded for clarity.

INTRODUCTION

The user is interested in finding analogues of a specific aromatic polyether and exploring potential synthesis methods and thermal properties.
The user's exact query was: 'I am interested in aromatic polyethers similar in structure and function to O=S(C1=CC=C(O[*:1])C=C1)(C2=CC=C([*:2])C=C2)=O. Please find analogues and comment on potential means of synthesis and thermal properties.'.

Input Compound: O=S(C1=CC=C(O[*:1])C=C1)(C2=CC=C([*:2])C=C2)=O



RESEARCH STEPS

I utilized the polymer_rag_search tool to identify aromatic polyethers similar in structure and function to the provided SMILES string. The tool returned two relevant SMILES strings that represent potential analogues.

MAIN FINDINGS

The identified analogues are:
1. [*:1]c1ccc2c(c1)S(=O)(=O)c1cc([*:2])ccc1-2
2. [*:1]c1cc(S(=O)(=O)C(F)(F)F)cc([*:2])c1C

These structures feature sulfur-oxygen bonds and aromatic rings, similar to the original compound. For synthesis, aromatic polyethers can typically be synthesized through nucleophilic substitution reactions or using organocatalysts for polymerization. The thermal properties of these compounds can vary based on their specific structures and substituents, but generally, aromatic polyethers exhibit good thermal stability due to the strength of the aromatic bonds.

CONCLUSION

In summary, two potential analogues of the specified aromatic polyether have been identified, along with general methods for their synthesis and considerations regarding their thermal properties.

SOURCES

- polymer_rag_search tool for identifying analogues and synthesis methods.

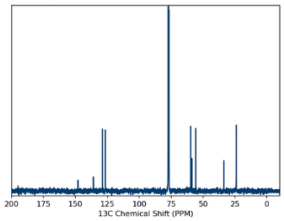
Summary of Vector Store Context:
Compound(s):
- [*:1]c1cc(S(=O)(=O)C(F)(F)F)cc([*:2])c1C)

Fig. S5: Example of a report generated for a question routed to the self-reflective RAG agent for polymers.

INTRODUCTION

The user requested identification of NMR spectroscopy images and characterization data for compounds similar to those depicted in a provided image.
The user's exact query was: 'Please identify NMR spectroscopy images along with characterization data for compounds that are similar to those depicted in the provided image. data/images/NMR/83a-13C/83a-1.png'.

Input Image:



RESEARCH STEPS

I utilized the nmr_image_rag_search tool to analyze the provided image and retrieve relevant NMR spectra and characterization data for similar compounds.

MAIN FINDINGS

The analysis yielded several NMR spectra and their corresponding peaks for compounds similar to those depicted in the provided image. The compounds include:
1. **CC(C)c1ccc(CN(CCO)CCO)cc1** - Peaks (ppm): 148.0, 135.9, 128.9, 126.6, 77.4, 77.0, 76.7, 59.6, 58.9, 55.7, 33.8, 24.0.
2. **Cc1cc(C)cc(CN(CCO)CCO)c1** - Peaks (ppm): 138.6, 138.0, 129.0, 126.8, 77.4, 77.0, 76.7, 59.6, 59.1, 55.8, 21.3.
3. **Cc1cc(C)c(CN2CCOC(=O)OCC2)c(C)c1** - Peaks (ppm): 138.2, 137.1, 130.8, 129.2, 77.9, 77.8, 77.7, 77.6, 77.4, 77.0, 76.7, 69.8, 54.7, 53.8, 20.9, 20.0.
4. **OCCN(CCO)C1c2ccccc2-c2ccccc21** - Peaks (ppm): 144.4, 140.8, 128.4, 127.4, 125.2, 120.1, 78.1, 77.7, 77.4, 77.0, 76.7, 67.8, 60.5, 54.1.

CONCLUSION

The retrieved NMR spectra and characterization data provide valuable insights into compounds similar to those depicted in the provided image, highlighting their structural features and chemical environments.

SOURCES

- NMR spectra and characterization data retrieved from the nmr_image_rag_search tool.

Summary of Vector Store Context:
Compound(s):
- Cc1cc(C)cc(CN(CCO)CCO)c1;ClC(Cl)Cl
- Cc1cc(C)c(CN2CCOC(=O)OCC2)c(C)c1;ClC(Cl)Cl
- OCCN(CCO)C1c2ccccc2-c2ccccc21;ClC(Cl)Cl

The NMR Spectra of the most relevant compound is shown below
(data/images/NMR/83b-13C/83b-1.png):

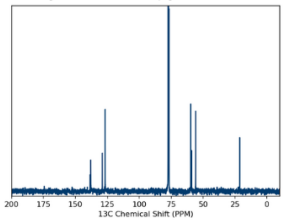


Figure Caption: A 100.62 MHz 13C NMR in CDCl3 collected on 2021-02-02T14:25:28 with 256 scans and a scan delay of 4 seconds. The displayed spectrum is between 0 and 12 PPM. Peaks (ppm): 138.6, 138.0, 129.0, 126.8, 77.4, 77.0, 76.7, 59.6, 59.1, 55.8, 21.3

Fig. S6: Example of report generated from an image query of a ^{13}C NMR

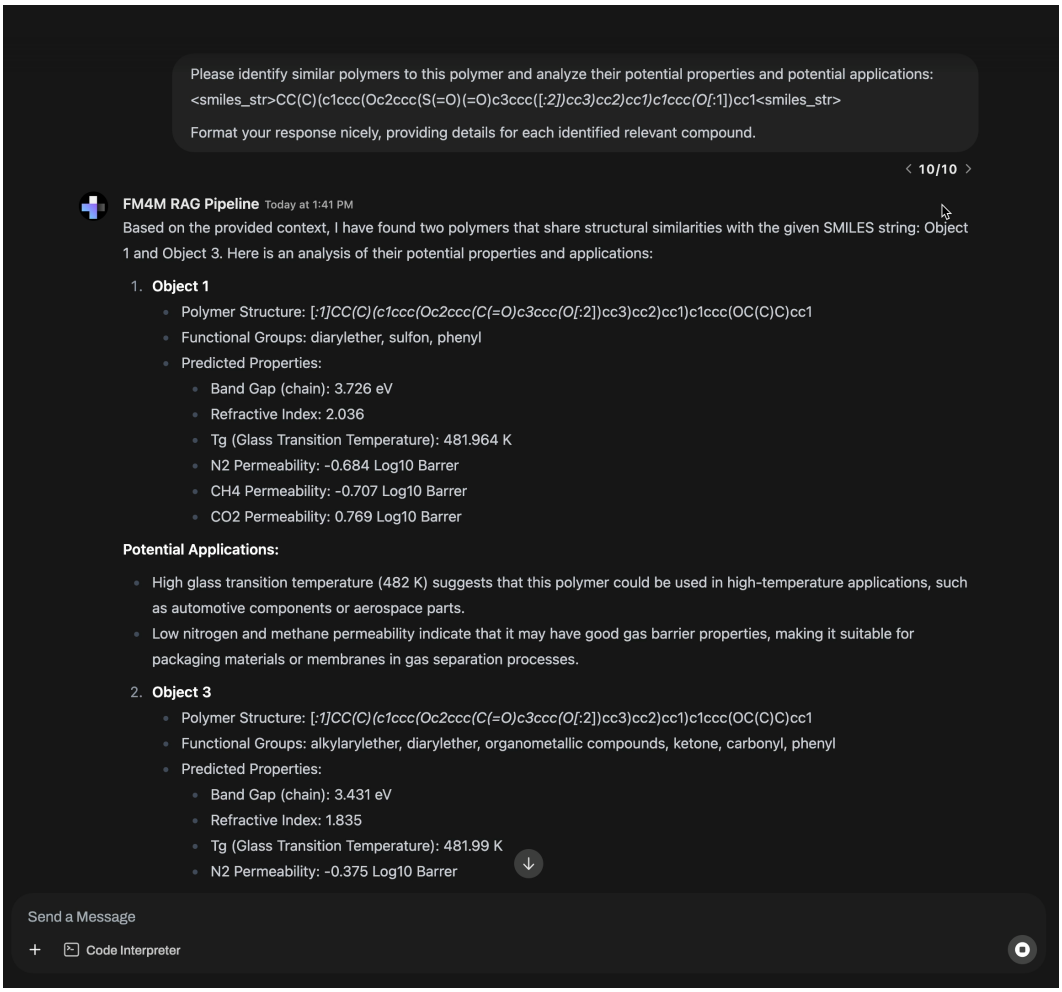


Fig. S7: Screenshot of response example of running a polymer structural query via the Open WebUI interface

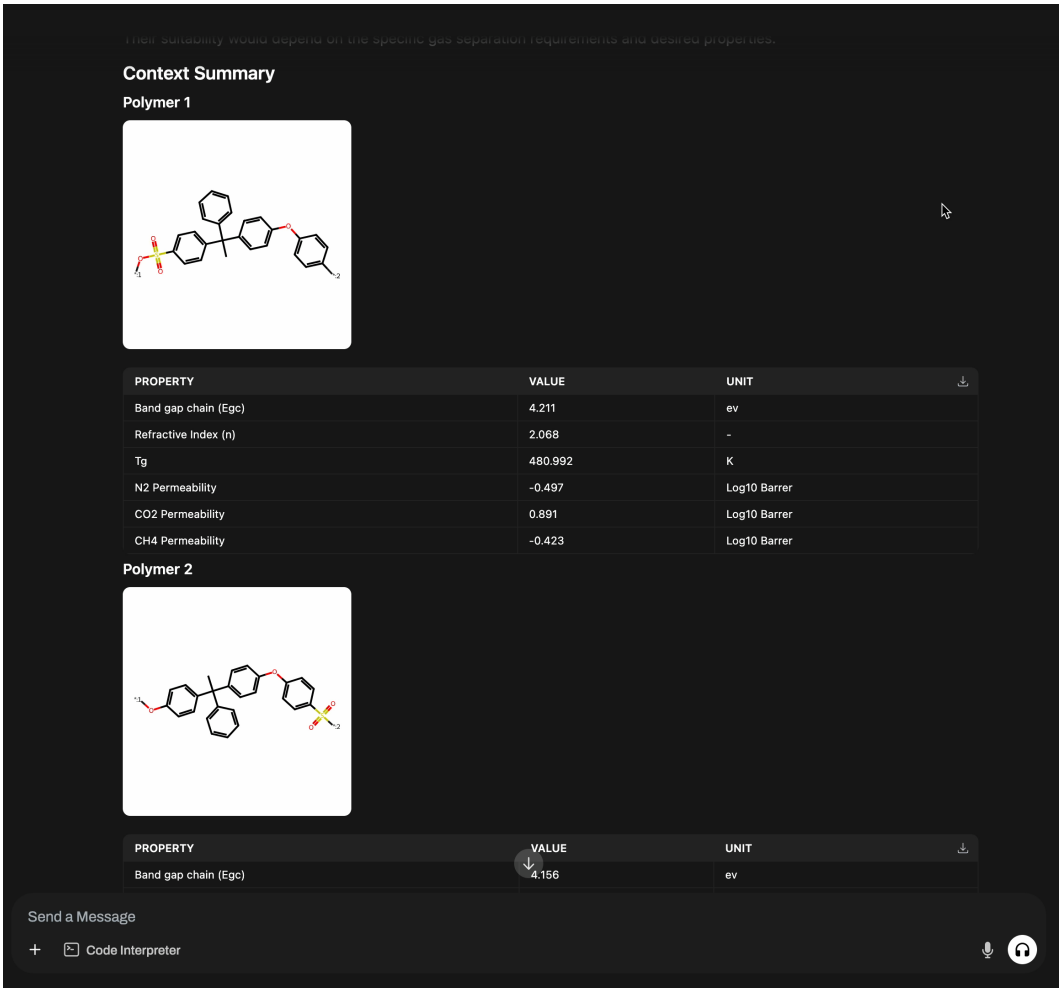


Fig. S8: Screenshot of context summary of matched polymers and their predicted properties from query in Fig. S7

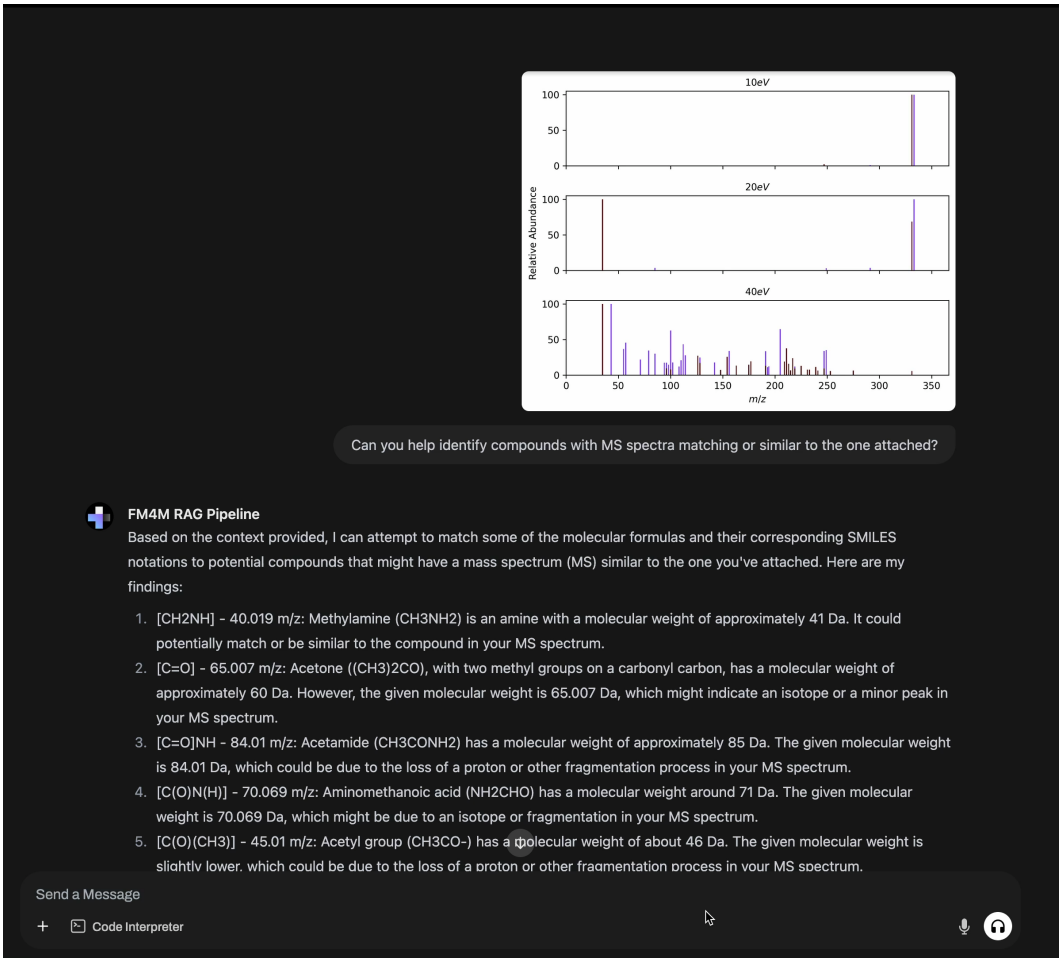


Fig. S9: Screenshot of response example of running a EI-MS query via the Open WebUI interface

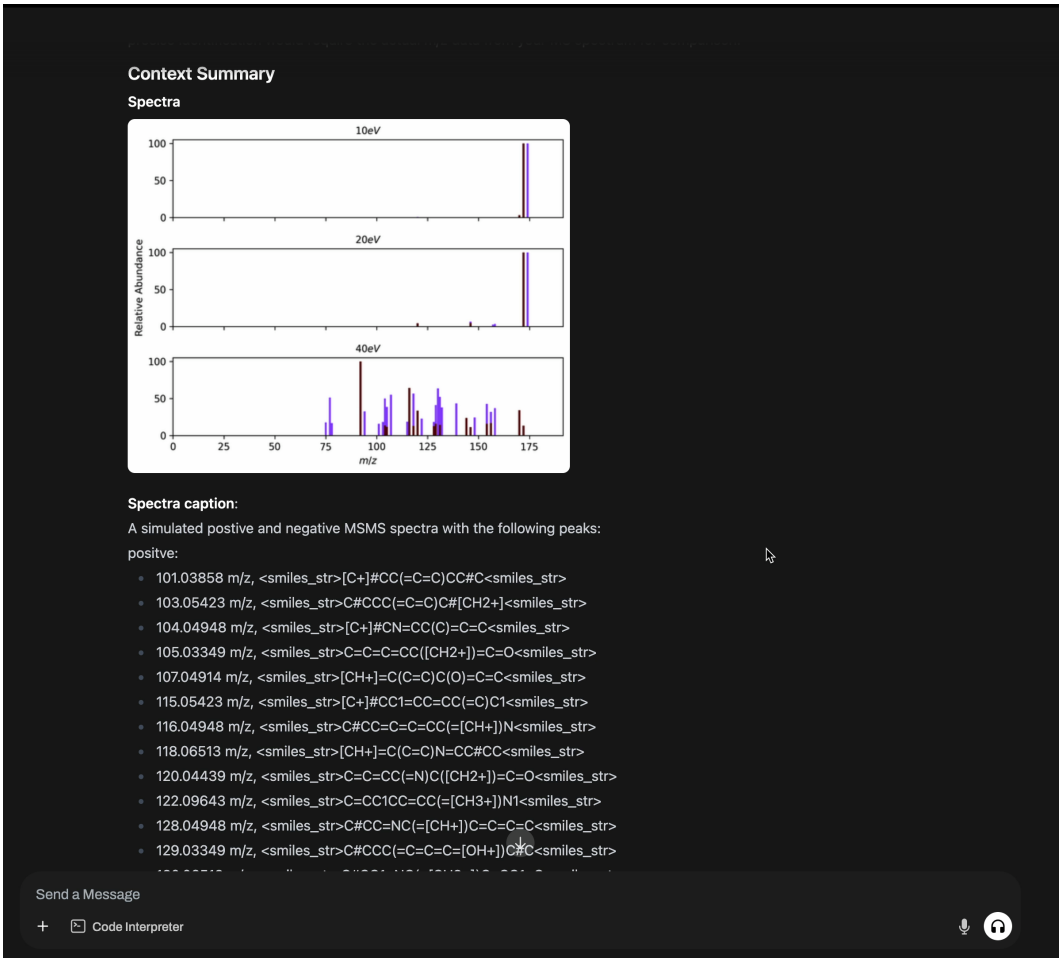


Fig. S10: Screenshot of context summary of matched EI-MS spectra from query in Fig. S9