

# Targeted Multi-Modal Passage Search for Molecules and their Synthesis Pathways

Abhisek Dey  
Rochester Institute of Technology  
Rochester, New York, USA  
ad4529@rit.edu

Nathaniel H. Stanley  
Insitro  
South San Francisco, California, USA  
nate@insitro.com

Richard Zanibbi  
Rochester Institute of Technology  
Rochester, New York, USA  
rxzvc@rit.edu

## Abstract

We present a chemical extraction and search pipeline intended to support information tasks related to drug discovery. Commonly used search tools for drug discovery such as Reaxys and SciFinder do not allow users to obtain retrieval results at the passage level. To address this, we present a passage retrieval tool for chemical patents that supports queries combining text and molecule diagrams expressed in SMILES. When SMILES is provided as a part of a query, the system refines text retrieval results through matching both textual names and drawn figures based on extracted SMILES representations. Molecule matches are obtained through substructure matching and structural similarity. This functionality was motivated by a chemist's need to find synthesis pathways for specific molecules containing a substructure of interest that binds and thus inhibits specific human genes. For this demonstration, we index a collection of 131 PDF patents categorized into 12 specific genes enabling a user to search on them. There are 32,301 document pages in the collection. Our user interface can be accessed at <https://unichemfinder.gccis.rit.edu/>. Our source code and data is available at <https://gitlab.com/dprl/unichemfinder>.

## CCS Concepts

• **Information systems** → **Search interfaces**; **Multimedia and multimodal retrieval**; **Information extraction**; **Document filtering**.

## Keywords

chemical multi-modal search, molecular diagram extraction, chemical named entity recognition, chemical passage search, chemical diagram search, multi-modal querying

## ACM Reference Format:

Abhisek Dey, Nathaniel H. Stanley, and Richard Zanibbi. 2025. Targeted Multi-Modal Passage Search for Molecules and their Synthesis Pathways. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25)*, July 13–18, 2025, Padua, Italy. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3726302.3730149>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).  
SIGIR '25, Padua, Italy

© 2025 Copyright held by the owner/author(s).  
ACM ISBN 979-8-4007-1592-1/2025/07  
<https://doi.org/10.1145/3726302.3730149>

## 1 Introduction

Specialized Chemical Information Retrieval (CIR) has traditionally been a difficult domain to make meaningful progress in within the open source community. The limitations stem from the majority of chemical research being available only in documents such as patents and publications, where information appears in different forms including text, molecule diagrams, synthesis pathway diagrams, tables and charts. Oftentimes, there are implicit links between the text and other forms of available information which is essential to get a complete understanding for a topic. Compounding the problem is the lack of a standardization in how information is represented by researchers. For example, the IUPAC [16] name *Ethyl Acetate* can be represented in other forms such as by the two chemical formulas  $C_4H_8O_2$  and  $CH_3CH_2CO_2CH_3$ , or abbreviated as *EtOAc*, *ETAC*, or *EA*. The diagram of this compound can also be represented in different ways. Furthermore, intellectual property constraints limit open research.

Attempts at CIR have included ChemXSeer [11], TREC-CHEM [8], Text2Mol [2], MoleculeSTM [6] and SureCEMmBL [13]. These systems focus on specific parts of the retrieval problem such as property prediction or molecule search given a description. Their datasets were limited, and based on standalone pre-curated molecules and did not contain reactions.

Closed source systems such as SciFinder [3] and Reaxys<sup>1</sup> attempt to mitigate the limitations of chemical standards by allowing users to search for single compounds as text or SMILES [12, 17] – a string based representation for a molecule within patents and publications. For e.g.,  $CCOC(=O)C$  is the SMILES representation ethyl acetate. Reaxys also curates reaction information from documents but from our usage of the tool, we found that their query formulation requirements for chemists are not very user friendly. Furthermore, indexed reactions often suffer from low recall as their semi-automated extraction system uses a human-in-the-loop form of extraction where the raw data is fact checked by a domain expert. Thereby, reactions are oftentimes lacking essential accompanying data such reagents or catalysts used, reaction conditions etc. The most important limitation of such systems is the index does not record the pages or passages where indexed data appears. This is especially relevant in the context of drug discovery as chemists want to find related molecule properties and different ways a specific substructure that inhibits a specific gene is synthesized.

We developed UniChemFinder to address limitations in existing CIR systems (see Figure 1). At a glance, distinguishing features of our system are:

<sup>1</sup><https://www.reaxys.com/>

**Figure 1: UniChemFinder page level results for the text-only query “difluoromethyl pyrimidine obtained with 400MHz NMR”. Returned passages are shown in the left panel, while diagrams linked to the selected first passage are shown at right in a list. The ‘Expand View’ buttons allow users to see the full page associated with a passage or diagram in a pop-up window.**

- (1) Fully automated extraction of molecule names and diagrams, and linking names in passages with matching molecule diagrams anywhere in the PDF
- (2) Ability to search PDFs at the page and passage level along with any linked molecules diagrams – either standalone or as a part of a synthesis pathway
- (3) Queries with multiple compound names along with specific reaction conditions are supported in text
- (4) Supports SMILES queries with missing groups or substructures in them. E.g.,  $*C(=O)C$  or  $C(=O)C$
- (5) Multi-modal queries can be used to search for specific molecules of interest containing a substructure (SMILES Query) with certain reaction conditions or catalyst use (Text Query)
- (6) Curated test collection of 131 patent PDFs categorized by their target gene

The following sections briefly describe our extraction and linking procedure. Then we focus our attention on the user interface, and finally provide an example use case.

## 2 Data Extraction, Parsing and Linking

Molecule regions are first identified by using YoloV8 [4] trained on detection data from [1] and then parsed into its corresponding graph using our improved version of MolScribe [14]. These graphs were then converted into their SMILES representation using

RDKit<sup>2</sup>. On the text side, passages and image caption regions are first located in page images using *layoutparser* [15] followed by OCR using PyTesseract [5]. As many passages are not chemically relevant, we use ChemDataExtractor2.0 [10] to identify passages with valid chemical entity names. Some entity names may not be IUPAC names or full compounds. We then use OPSIN [7] to convert IUPAC names to SMILES. SMILES provides a common representation for molecular structure, which we use to link chemical entities in passages to equivalent diagrams. Passages may contain multiple molecule names, and each may be linked to one or more instances of the same drawn molecule.

For this paper, we extract and index chemical information from 131 PDF patents collected from publicly available patent organizations like USPTO<sup>3</sup> and WIPO<sup>4</sup>. This also served our purpose of making a search system catering specifically to the information needs of drug discovery. The patents were carefully curated based on the specific genes that the molecules in them targeted. This work focused on 12 main genes — CD73, DGAT2, DKGa, GLP-1, Helios, IRAK4, KHK, KRAS, Mcl1, PARP7, PRMT5 and TEAD. The total number of indexed pages is 32,301 with about 110k passages containing a valid IUPAC name. This helped solidify our understanding of “relevance” for small molecule drug discovery. Searching

<sup>2</sup>RDKit: Open-source cheminformatics. <https://www.rdkit.org>.  
<https://doi.org/10.5281/zenodo.591637>

<sup>3</sup><https://www.uspto.gov/>

<sup>4</sup><https://www.wipo.int/portal/en/index.html>

Passage Results

(0) Rank: 1, Gene: CD73, PDF: WO2024006929A1.pdf, Page: 97

Expand View


[0199] Representative synthetic Scheme 1 shows a general synthesis of compounds of the disclosure. The methodology is compatible with a wide variety of functionalities. In Representative Synthesis 1, a suitably substituted chloropyrimidine, chloropyridazine, or chloropyridine (or the corresponding bromo- or iodo- compound) is combined with a suitably substituted pyrrolidine in a suitable solvent system (e.g. *tert*-butanol, DMAc, dioxane, etc.) in the presence of a palladium catalyst (e.g. RuPhos Pd G3, Pd(OAc)<sub>2</sub> + XantPhos, etc.) and base (e.g. Cs<sub>2</sub>CO<sub>3</sub>, K<sub>3</sub>PO<sub>4</sub>, etc.) at elevated temperature (e.g. ranging from about 80 – 120 °C). Subsequently, the resultant suitably substituted 2,4-dimethoxypyrimidine-containing compound can be treated with an acid (e.g. hydrochloric acid) in a suitable solvent system (e.g. water + methanol) at elevated temperature (e.g. ranging from about 60 – 80 °C).

Passage Results

(1) Rank: 2, Gene: IRAK4, PDF: WO2014074657A1.pdf, Page: 196

Expand View

15




[00308] Synthesis of 5-chloropyrazolo[1,5-a]pyrimidine: A mixture of pyrazolo[1,5-a]pyrimidin-5-ol (60 g, 444 mmol) in acetonitrile (180 mL) in a 1L single neck RBF fitted with reflux condenser under magnetic stirring with N<sub>2</sub> outlet was added POCl<sub>3</sub> (112 mL, 1202 mmol) and then heated at 80 °C for 3h. The mixture was quenched into an ice cold solution of saturated NaHCO<sub>3</sub> slowly until pH =7-8 and then extracted in EtOAc (700 mL). The organic layer was separated and then washed with NaHCO<sub>3</sub> solution, followed by brine. The organic layer obtained was washed with 10% NaHCO<sub>3</sub> solution (150 mL) dried over Na<sub>2</sub>SO<sub>4</sub>, filtered and evaporated to give 5-chloropyrazolo[1,5-

Passage Results

(1) Rank: 1, Gene: IRAK4, PDF: WO2014074657A1.pdf, Page: 196

Expand View

15



[00308] Synthesis of 5-chloropyrazolo[1,5-a]pyrimidine: A mixture of pyrazolo[1,5-a]pyrimidin-5-ol (60 g, 444 mmol) in acetonitrile (180 mL) in a 1L single neck RBF fitted with reflux condenser under magnetic stirring with N<sub>2</sub> outlet was added POCl<sub>3</sub> (112 mL, 1202 mmol) and then heated at 80 °C for 3h. The mixture was quenched into an ice cold solution of saturated NaHCO<sub>3</sub> slowly until pH =7-8 and then extracted in EtOAc (700 mL). The organic layer was separated and then washed with NaHCO<sub>3</sub> solution, followed by brine. The organic layer obtained was washed with 10% NaHCO<sub>3</sub> solution (150 mL) dried over Na<sub>2</sub>SO<sub>4</sub>, filtered and evaporated to give 5-chloropyrazolo[1,5-

Passage Results

(0) Rank: 2, Gene: TEAD, PDF: WO2023150619A2.pdf, Page: 95

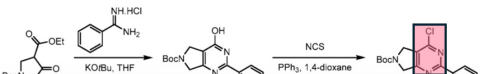
Expand View

5

tetramethyl-1,3,2-dioxaborolan-2-yl)-2H-1,2,3-triazole (R2) as a white solide (147 mg, 0.7 mmol, 57%). The product was used without further purification.

5

[00308] Synthesis of 4-chloro-2-phenyl-5,7-dihydropyrrolo[3,4-d]pyrimidine-6-carboxylate (int-1)



**Figure 2: Left: Search Result from text-only query "synthesis of flourooxetan pyrimidine". Right: Result for multi-modal query combining the text with the SMILES string "ClC1=NC=CC=N1". Right: The highlighted substructures are matches for the SMILES part of the query. This demonstrates how a multi-modal query helps in refining text results by re-ranking hits with a substructure provided as SMILES, mentioned in textual form in the passages.**

for specific compounds or reaction pathways should ideally return candidates belonging to a PDF for the targeted gene. This is inherently due to the fact that in drug discovery, specific substructures of interest are common in patents addressing a particular gene, differing primarily by their auxiliary atom groups.

An important distinction between other CIR systems and UniChemFinder is the text tokenization scheme. This enables users to be as constrained or as flexible as they like while formulating text queries for molecule searches. For example, complex IUPAC names like 6-(3-methoxyphenyl)quinazolin-4-amine are broken down into sub-groups "6 3 methoxy phenyl quinazoline 4 amine". This is done for both queries and indexed passages. Tokenization ensures that users who either do not exactly know the full molecule they are looking for or want to get molecules containing specific subgroups can do so. This is especially relevant for drug discovery, as chemists are often interested in a specific sub-group or a combination of sub-groups mentioned in the same passage. UniChemFinder thus offers a more sophisticated way of searching compound name mentions.

### 3 Modes of Search and Scoring

UniChemFinder supports three search modes: (1) text search using BM25 as implemented in PyTerrier [9], (2) individual molecular structures provided as SMILES are searched using RDKit's Tanimoto similarity search, and (3) a multi-modal search combining BM25 text search with molecular *substructure* search.

For searching on individual molecules given as SMILES, we use the Tanimoto similarity for molecules represented as sparse bit vectors. These 'fingerprint' bit vectors represent the presence or absence of atoms/subgroups when traversing a molecular graph. We use Morgan Fingerprinting as a specialized form of this traversal. The Tanimoto similarity is the number of common bits divided by the number of unique bits across the two vectors. In this way, the Tanimoto simailty is a form of Jaccard index or *intersection-over-union* (IoU) measure.

In contrast, the molecular *substructure* matching used for multi-modal search identifies molecules containing the SMILES query graph as a subgraph. In this way, our substructure search is a form of boolean filter: all molecules in the index containing the graph corresponding to a SMILES query are returned.

For multi-modal search, a SMILES query is matched to substructures in molecules appearing as IUPAC names in passages. The final multi-modal ranking is a modified conjunctive search for keywords and molecular substructures: retrieved text passages are re-ranked using a two-level sort, ordered by decreasing number of substructure matches for a passage and then decreasing BM25 score. Passages matching query SMILES but not query text are omitted. By treating text passages as the grounding modality, we address information needs that benefit from the additional constraints introduced by molecular structure matching, as seen in Figure 2. This also avoids empty results, which is a common occurrence in other CIR systems when searching for specific compounds or reactions.

## 4 User Interface

Figure 1 shows the interface for our system. The interface exposes two search boxes in the top navigation bar, with the left box for text queries and the right box for SMILES queries. When the search button is pressed, the search mode is selected based on which query input boxes are empty. Text-search mode is used when the SMILES box is empty and vice versa. Multi-modal search executes when both text and SMILES are provided. Users can get acclimated to the system by following the instructions provided in a pop-up window after clicking the *How To* button at top right.

After a search is executed, the left panel shows the retrieved passages. This compact view shows a fixed-size window around matched passages where they appear in a PDF document. This enables users to quickly skim through the results to find passages of interest to them in-context. Each retrieved passage is also accompanied by metadata including the page number, PDF Filename, and Target Gene. The first number in the passage banners is the number of linked diagram matches for a passage. This makes it easier for the user to locate passages containing the more or fewer diagram links without having to individually click on passages.

The right panel shows molecules in diagrams linked to the currently selected passage. Whenever a passage from the left panel is clicked, that passage is highlighted, and if it has linked diagrams associated with it, they are displayed as a list of windowed PDF page views in the right panel. This is demonstrated in Figure 1, where the first passage at left has been clicked. This is particularly useful, as a single molecule may be drawn at multiple places in the PDF. This allows a user to quickly navigate through pages of interest to find places where the molecule has been used in a very specific context. Passages can also be linked to more than one molecule; the metadata banners for diagrams in the right panel of Figure 1 include IUPAC names and SMILES for matched compounds, to help users quickly identify different molecules. Page numbers are also provided in the banners. Furthermore, including SMILES with diagram matches is helpful for cheminformatics, as SMILES can be directly used in downstream tasks such as tools for molecule modeling and property prediction directly.

The expanded view window opens when a user wants additional information about a passage or a diagram match. Users can switch from a compact view to an expanded view showing the full page of the selected passage or a diagram in one place. This mode can be activated by clicking on the "Expand View" button attached to each hit. Our tool tracks the hit of interest and provides a user the option to cycle through all the other diagram hit pages for a particular passage in an interactive way.

## 5 Use-Case: Search for Pyrimidine Compounds

Pyrimidine compounds are particular interest to chemists in drug discovery, as they are one of the building blocks of DNA. It has been found to be useful in different forms for the synthesis of medicines that inhibit different genes in humans such as CD73, IRAK4 and TEAD. These genes have been known to cause debilitating diseases in humans. However, in a chemical search context, the challenge arises from the fact it is used in vastly different gene targeting molecules along with different auxiliary atom groups. This essentially means that a simple text search for the compound might

not yield relevant results if the user wants to find highly specific molecular information for a particular gene or carrying some other specific atom groups.

Figure 2 shows the comparison of search between a text-only and a multi-modal query for synthesis information of fluorooxetan pyrimidine, and a more specific search by including a substructure of interest, C1C1=NC=CC=N1. We see in this case that using the text only mode results in a broader hit of pyrimidine compounds for the first hit like chloropyridazine. However, on specifying a substructure of interest which contains a chlorine atom in the pyrimidine compound, the search results refer to passages that are more likely to contain references to them. Furthermore, the images around the vicinity of the passage hits were also more likely to refer to compounds containing the substructure of interest as shown by the pink highlights in Figure 2. This is a particularly challenging search, as there are many gene inhibitors that use pyridine compounds, sometimes being different by only a single substituent atom. It is worth noting that the search results included passages containing the terms "fluoro" and "oxetan" in later ranks individually. However, as BM25 weighs additional text token matches higher, any passage with the terms "synthesis of" were automatically ranked higher. We will address this behavior with an improved model in future. This instance shows how important and useful multi-modal queries can be in certain scenarios.

## 6 Conclusion

We have presented UniChemFinder, our first attempt at fusing queries in different modalities in the chemical domain for passage search in PDF documents. Our interface exposes this functionality through three search modes that can be flexibly used to suit a user's/chemist's information need. Our compact view of search results directly inside source PDFs were motivated by chemists' need to find related information not directly attached to a chemical query – be it text or diagrams.

We also developed a novel linking method through which passages containing molecule name mentions were directly linked to drawn molecule diagrams. Our work includes a curated collection of patent PDFs that users search. We plan to have an end-to-end extraction, indexing and search system in the future where a user can provide their own collection of PDFs for indexing and search. UniChemFinder serves as a baseline system for future work where we anticipate using passages and SMILES together for dense search. Ideally, we would also like to implement a third modality of search where users can search through image snippets of molecules and reaction pathways. We believe that such a retrieval approach would be useful for Retrieval Augmented Generation (RAG).

## Acknowledgments

This work was supported by the National Science Foundation (USA, Grant #2019897, Molecule Maker Lab Institute). We would also like to thank Kent Gorday from Insitro who assisted in manually compiling the patents for the test collection.

## References

- [1] Abhisek Dey and Richard Zanibbi. 2021. *ScanSSD-XYc: Faster Detection for Math Formulas*. Vol. 12916 LNCS. Springer International Publishing. 91–96 pages. doi:10.1007/978-3-030-86198-8\_7

- [2] Carl Edwards, Cheng Xiang Zhai, and Heng Ji. 2021. Text2Mol: Cross-Modal Molecule Retrieval with Natural Language Queries. *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings* (2021), 595–607. doi:10.18653/v1/2021.emnlp-main.47
- [3] Stephen Walter Gabrielson. 2018. SciFinder. 588-590 pages. Issue 4. doi:10.5195/JMLA.2018.515
- [4] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. 2023. *Ultralytics YOLOv8*. <https://github.com/ultralytics/ultralytics>
- [5] Anthony Kay. 2007. Tesseract: an open-source optical character recognition engine. *Linux J.* 2007, 159 (July 2007), 2.
- [6] Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei Xiao, and Anima Anandkumar. 2022. Multi-modal Molecule Structure-text Model for Text-based Retrieval and Editing. (2022), 1–31. <http://arxiv.org/abs/2212.10789>
- [7] Daniel M. Lowe, Peter T. Corbett, Peter Murray-Rust, and Robert C. Glen. 2011. Chemical Name to Structure: OPSIN, an Open Source Solution. *Journal of Chemical Information and Modeling* 51, 3 (2011), 739–753. doi:10.1021/ci100384d arXiv:<https://doi.org/10.1021/ci100384d> PMID: 21384929.
- [8] Mihai Lupu, Jimmy Huang, Jianhan Zhu, and John Tait. 2009. TREC-CHEM: large scale chemical information retrieval evaluation at TREC. *SIGIR Forum* 43, 2 (Dec. 2009), 63–70. doi:10.1145/1670564.1670576
- [9] Craig Macdonald, Nicola Tonellotto, Sean MacAvaney, and Iadh Ounis. 2021. PyTerrier: Declarative Experimentation in Python from BM25 to Dense Retrieval. In *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021*, Gianluca Demartini, Guido Zuccon, J. Shane Culpepper, Zi Huang, and Hanghang Tong (Eds.). ACM, 4526–4533. doi:10.1145/3459637.3482013
- [10] Juraj Mavračić, Callum J. Court, Taketomo Isazawa, Stephen R. Elliott, and Jacqueline M. Cole. 2021. ChemDataExtractor 2.0: Autopopulated Ontologies for Materials Science. *Journal of Chemical Information and Modeling* 61 (9 2021), 4280–4289. Issue 9. doi:10.1021/acs.jcim.1c00446
- [11] Prasenjit Mitra, C. Lee Giles, Bingjun Sun, and Ying Liu. 2007. ChemXSeer: a digital library and data repository for chemical kinetics. In *Proceedings of the ACM First Workshop on CyberInfrastructure: Information Management in ESience* (Lisbon, Portugal) (*CIMS '07*). Association for Computing Machinery, New York, NY, USA, 7–10. doi:10.1145/1317353.1317356
- [12] Noel O'Boyle and Andrew Dalke. 2018. DeepSMILES: An Adaptation of SMILES for Use in Machine-Learning of Chemical Structures. *ChemRxiv* (2018), 1–9. doi:10.26434/chemrxiv.7097960
- [13] George Papadatos, Mark Davies, Nathan Dedman, Jon Chambers, Anna Gaulton, James Siddle, Richard Koks, Sean A. Irvine, Joe Pettersson, Nicko Goncharoff, Anne Hersey, and John P. Overington. 2016. SureChEMBL: A large-scale, chemically annotated patent document database. *Nucleic Acids Research* 44 (2016), D1220–D1228. Issue D1. doi:10.1093/nar/gkv1253
- [14] Yujie Qian, Jiang Guo, Zhengkai Tu, Zhening Li, Connor W. Coley, and Regina Barzilay. 2023. MolScribe: Robust Molecular Structure Recognition with Image-to-Graph Generation. *Journal of Chemical Information and Modeling* 63 (2023), 1925–1934. Issue 7. doi:10.1021/acs.jcim.2c01480
- [15] Zejiang Shen, Ruochen Zhang, Melissa Dell, Benjamin Charles Germain Lee, Jacob Carlson, and Weining Li. 2021. LayoutParser: A Unified Toolkit for Deep Learning Based Document Image Analysis. *arXiv preprint arXiv:2103.15348* (2021).
- [16] Stanislaw Skonieczny. 2006. The IUPAC rules for naming organic molecules. *Journal of Chemical Education* 83 (2006), 1633–1637. Issue 11. doi:10.1021/ed083p1633
- [17] David Weininger. 1988. SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules. *Journal of Chemical Information and Computer Sciences* 28 (1988), 31–36. Issue 1. doi:10.1021/ci00057a005