In-Context Retrieval for Molecules and Chemical Synthesis Pathways

by

Abhisek Dey

A dissertation submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
in Computing and Information Sciences

B. Thomas Golisano College of Computing and
Information Sciences

Rochester Institute of Technology
Rochester, New York
April 2026

Signature of the Author _____

Certified by _____
PhD Program Director Date

In-Context Retrieval for Molecules and Chemical Synthesis Pathways

by

Abhisek Dey

Committee Approval:

We, the undersigned committee members, certify that we have advised and/or supervised the candidate on the work described in this dissertation. We further certify that we have reviewed the dissertation manuscript and approve it in partial fulfillment of the requirements of the degree of Doctor of Philosophy in Computing and Information Sciences.

Dr. Richard Zanibbi — Date
Dissertation Advisor

Dr. Nathaniel H. Stanley — Date
Dissertation Committee Member

Dr. Weijie Zhao — Date
Dissertation Committee Member

Dr. Matthew Fluet — Date
Dissertation Committee Member

Dr. Dan Phillips — Date
Dissertation Defense Chairperson

Certified by:

[Dr. Pengcheng Shi] — Date
Ph.D. Program Director, Computing and Information Sciences

In-Context Retrieval for Molecules and Chemical Synthesis Pathways

by

Abhisek Dey

Submitted to the
B. Thomas Golisano College of Computing and Information Sciences Ph.D. Program in Computing and Information Sciences
in partial fulfillment of the requirements for the
Doctor of Philosophy Degree
at the Rochester Institute of Technology

Abstract

Contrastive learning methods require well-defined positive pairs, limiting their applicability to domains where complete, high-fidelity pairings are available. In practice, large-scale scientific corpora –including patents, publications, and web-scale data – contain vast quantities of contextually relevant but incompletely paired samples that are discarded under standard training paradigms. In this work, we demonstrate that hard negative mining can be leveraged to construct pseudo-positive supervision signals from unpaired or partially paired data, enabling contrastive learning to exploit the full breadth of available corpora without sacrificing representational quality. Using a large-scale chemical drug patent corpus as a testbed, we train a cross-modal contrastive model aligning chemical text passages with molecular SMILES representations, where a significant fraction of passages lack extractable chemical entity mentions, and have no ground-truth positive pairing. We show that incorporating these unpaired passages via hard-negative-derived pseudo-positive objectives yields higher ranking quality (nDCG) compared to training on strictly paired data alone, with the hard negative selection strategy proving critical – random pseudo-positive assignment degrades performance while geometric hard negative mining provides meaningful alignment signal. Our findings suggest that partially paired data, when coupled with principled pseudo-supervision, provides complementary context that enriches the contrastive learning objective beyond what high-fidelity pairs alone can offer. This paradigm generalizes naturally to other data-abundant but incompletely paired domains, including biology, physics, and large-scale web and social media corpora, offering a path toward more data-efficient contrastive representation learning at scale.

To rigorously evaluate our model, we constructed an end-to-end test collection grounded in real-world chemical information retrieval needs, First, in collaboration with domain expert chemists, we developed a graded relevance benchmark comprising 35 diverse query topics, each with 10–30 expert-assessed candidates pooled from a retrieval pipeline combining specialized models for both

chemical text and SMILES-based molecular search. Relevance judgments were assigned at multiple granularity levels reflecting the degree to which each candidate satisfies the query's intended information need, yielding the first publicly available graded relevance collection for chemical information retrieval. Second, we constructed a multi-modal test dataset derived from 142 chemical patent PDFs spanning over 30,000 pages, covering compounds associated with 14 specific human gene targets. Each extracted passage retains provenance metadata linking it to its source PDF, page, and location, enabling fine-grained retrieval evaluation. The pooling pipeline used to generate candidates combined term-based retrieval (BM25, PL2), BERT-based dense retrieval for semantic text matching, Tanimoto similarity and subgraph search for molecular structure matching, and ChemBERTa-based dense retrieval for molecular semantic similarity. Multi-modal queries combining text and molecular inputs were resolved via overlap-based re-ranking over the union of text and molecule retrieval candidates. Together, the collection provides a rigorous and reproducible evaluation framework for cross-modal chemical retrieval.

Acknowledgments

I would like to firstly thank my advisor Dr. Richard Zanibbi for his mentorship and technical guidance throughout my PhD for making me a better researcher and encouraging me to dive into challenging problems and try to come up with unique solutions. Although in many occasions, I was not successful, the learning process made it worthwhile and made me learn the value of breaking down complex problems into very small pieces.

I would like to specially thank my manager at Insitro, Nate Stanley, who gave me the academic freedom and chemistry guidance to develop a better and more reliable version of the chemical diagram parser, MolScribeV2. What started as a summer intern project quickly developed into a mature system and has been used for projects concerning chemical entity linking as well as extraction for indexing chemical documents in a retrieval setting. Other notable mentions at Insitro are Patrick Conrad, Srinivasan Sivanandan, Kent Gorday, Tomasz Danel and Benson Chen for their insightful advice and infrastructure help for the project.

The past and present members of the Document and Pattern Recognition Lab (DPRL), RIT have been instrumental in their support in and outside of the lab. Special mention to Behrooz Mansouri who helped me settle in the DPRL. Other notable mentions include Ayush Kumar Shah, Bryan Amador and Patrick Phllipy.

I am also thankful to the Molecule Maker Lab Institute (MMLI) and National Center for Supercomputing Applications (NCSA) at University of Illinois at Urbana-Champaign (UIUC) for giving me the opportunity to develop my research skills in a cross-domain discipline of AI and chemistry. Special mention to Blake Ocampo and David Bianchi who have been great colleagues in helping me to diving into the intricacies and challenges of chemical AI and providing me with enough compute resources when it was needed the most.

This work was made possible through the support of National Science Foundation (USA) under Grant Nos. IIS-1717997 (MathSeer project) and 2019897 (Molecule Maker Lab Institute project). I am deeply appreciative of this support, which has facilitated my research and collaborations.

Special thanks to my folks, Suchira Dey and Subhasis Dey who have been there through thick and thin. Even though we don't see each other for years at a time and they being continents away, they have always supported my academic journey for which I will be forever indebted to them.

Finally, I would like to express my gratitude to mentors, and collaborators who have, directly or indirectly, supported my research endeavors.

Co-Authorship

Chapter 3 includes work on the chemical diagram parser which was initially coded at the Company Insitro Inc. during my time as an Summer Intern in 2024 under the guidance of Dr. Nathaniel Stanley and is appearing as a paper in International Joint Conference on Artifical Intelligence (IJCAI) 2025. It also contains the extraction and linking work, the development of which was guided by Dr. Richard Zanibbi.

Chapter 4 describes a test collection for evaluating the developed model. The original PDFs for the generated index were curated by Dr. Nathaniel Stanley and Dr. Kent Gorday. The sparse retrieval models and the user interface were developed under the guidance of Dr. Richard Zanibbi. Both of these are appearing in two papers in the Special Interest Group in Information Retrieval (SIGIR) 2025.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In Chemical Information Retrieval (CIR), specifically CIR for drug discovery, chemists have very specific information needs. More often than not, chemists focus on gathering information about specific inhibitor molecules from a subset of patents that inhibit a specific gene in humans. They look for high potential molecules judging by their synthesis pathways, intermediate reactants, products, substrates and reaction conditions. In addition, due to the vast number of closely related potential molecules described in each patent document, they shortlist viable molecules based on their activity, pharmacokinetics and pharmacogenetics data also available in the documents. However, the entire process of manually compiling the relevant gene inhibitor documents (e.g. GLP-1 (Ozempic), DGAT2, CD73 etc.), searching the documents for relevant molecules and their associated data is tedious and time consuming.

1.1 Extraction and Retrieval in Cheminformatics

Commonly used chemical search tools like SciFinder [41] and PubChem [60] do not effectively address this information need. The biggest hindrance of these systems is the need for inspection by trained chemists to manually curate or fact check the compiled database at some stages of the indexing process. For instance, SciFinder indexes only chemical compound information from text blocks available in patents and papers. Although they allow users to search diagrams by explicitly drawing them out in their online drawing tool, a lot of information available through drawn 2D molecule diagrams and reactions is lost in the process. In addition, they do not support multi-compound search in a single text-based query to further fine-tune the results. Using the

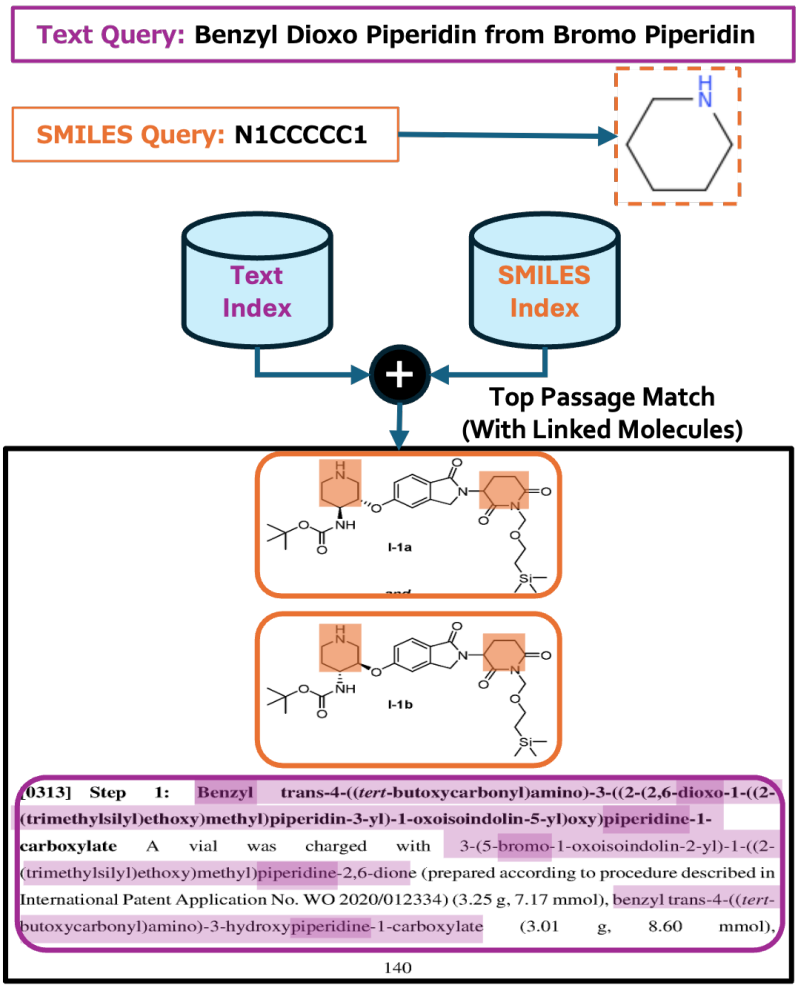

Figure 1.1: Given a query combining text tokens (maroon) and molecular sub-structure(s) in SMILES, the system combines text and SMILES matches to produce final fused hits by re-ranking hits from the two modes of search based on overlaps to produce a final ranking of PDF patent pages. The passage outlined in maroon along with linked molecule diagrams outlined in orange show the top matched result. Highlighting shows matches for text tokens (maroon) and matching molecular substructures (orange). The lighter maroon highlights refers to the full chemical entity match while the darker highlights refer to the text tokens that matched with the passage tokens.

patent/article search section in SciFinder using a single compound query points the users to only the likely PDFs that might contain the query molecule. A user needs to manually inspect the PDF to find the specific molecule and its associated data.

This work aims to addresses the challenges described above. Figure 1.1 shows how search works

for our system [33]. Given a text query containing more than one compound mentioned in it correct opening quotes ("Benzyl Dioxo Piperidin" & "Bromo Piperidin") and a substructure of interest in the form of a Simplified Molecular Input Line Entry System (SMILES) [147] string (`N1CCCCC1`) the system tokenizes the text and breaks down the molecule names into sub-groups (e.g., 1-3 dinitrobenzene into [1, 3, di, nitro, benzene]) and searches the text index. Concurrently, the SMILES string referring to a structure is used to search the SMILES index using the RDKit[1] to find matching SMILES. The SMILES matches are linked to the PDF passages containing them. The combined result for text and SMILES search is shown to the user at the passage level in the PDF along with the matching diagrams.

In our work, search features are made possible due to automated (1) text extraction along with Chemical Named Entity Recognition (CNER), (2) molecule diagram detection and parsing into SMILES, and (3) linking chemical names from passages with their associated drawn molecules. In contrast to SciFinder, text queries can also search for specific reaction conditions (e.g. "reaction at 400 MHz NMR") or pharmacokinetics (e.g., "half-life $T_{\frac{1}{2}} = 15hrs$" or "follows metabolic pathway CYP450"). Also, most chemists are not entirely certain about the exact compound they are looking for and use only a part of the molecule as a query (e.g., in Figure 1.1, Bromo Piperidin is only a part of a variety of possible compounds with those sub-groups). These results are produced using a manually curated collection of 142 PDFs relating to 14 different genes and associated queries with graded relevances.

The next section will give an overview of the research questions and technical improvements considered for this proposal.

1.2 Research Questions

This work demonstrates the early foundational work to enable better in-context navigation and search for chemical molecules and synthesis pathways. Building on this, the later stages of this manuscript talks about the creation process of an expert curated test collection and defining a novel objective function to use ungrounded text passages to train dense retrieval models that are evaluated using this collection. We will delineate the research questions into ones that have been addressed in preliminary work and the subsequent research questions following the initial work.

[1]RDKit: Open-source cheminformatics. https://www.rdkit.org. https://doi.org/10.5281/zenodo.591637

1.2.1 Research Questions Addressed in Preliminary Work

1. **How to enable in-context search of chemical passages and associated diagrams in PDF using text based reaction conditions or molecule functional groups or SMILES structures (full or partial or both?)** As described in Section 3.3, we link all chemical IUPAC [132, 64] names (e.g. 1,3,5-trinitrobenzene) within indexed passages to their diagrams anywhere in the PDF. This was made possible through converting text-based IUPAC names and molecule diagrams into a common representation called SMILES [147]. This enabled passages to be linked to diagrams, even when they were hundreds of pages apart. As described in Section 3.1, we use a pipeline to extract text from PDF documents through PDF-to-Image conversion, Text block detection and finally OCR on the text-blocks to parse the text. The extracted text is then sent through a Chemical Named Entity (CNER) system to detect IUPAC names using string matching which are eventually converted to SMILES.

2. **Is there a better way to fuse Text + SMILES queries without using boolean models found in existing tools?** As described in Section 4.3, we use a two level sorting mechanism, where the first sort is based on the number overlaps in passages returned from the two independent modes of search. The second stage re-ranks text search candidates without overlap between the SMILES candidates by order of their original BM25 scores.

3. **How can we create a test collection with meaningful relevance criteria for patent documents containing chemical entities in text and drawn figures?** As described in Section 4.4, we design a test collection from our generated multi-modal index of 131 chemical PDF patents containing 57k passages from 32k pages. To evaluate the test collection, we designate relevance levels to each created topic by using the corresponding Gene, PDF and exact passage as relevance levels. This allows an exact-item retrieval task to have meaningful relevance ratings without manual assessment.

1.2.2 Technical Improvement Goals & Research Questions

Technical Improvement Goals

T1 Improve the number of detected passages to increasing the test collection size: As described in Section 5.1, we use a neural passage detection and OCR system to replace the currently used LayoutParser detection and Tesseract OCR. This neural-based end-to-end detection + OCR system greatly improved recall by indexing more chemically relevant

passages that are complete without missing text lines.

T2 Increase the number of passages linked to diagrams: As described in Section 5.1, we significantly increased the number of valid chemical compounds converted to SMILES by supporting common and industrial names and obtaining the associated SMILES to a chemical name. For e.g., if "1,3-Dinitrobenzene" is referred to in a PDF by its industrial name "AI3-02913", it can now be identified and converted to its correct SMILES.

T3 Search the test collection for chemical information about molecules and reactions from both text and drawn diagrams As described in Section 5.2.2, we create a novel pooling technique where we index and search for both chemical passages and diagrams, prioritizing the hits where a candidate contains both the passage and an associated molecule diagram.

Research Questions

RQ1 How to ensure that the test collection contains diverse multi-modal queries and relevance criteria for an exact match is factually relevant to a domain expert in the context of chemical search in drug discovery? As described in Section 5.2, we address the query generation procedure followed before in preliminary work to make it more chemically relevant and fitting into the information needs of chemists in drug discovery. This included drawing in chemical experts to create queries that reflected real-world complex information needs. The same queries were scored by the chemists to create a diverse relevance test collection. The queries were multi-modal in nature.

RQ2 How can multi-modal self-supervised contrastive learning be used to align SMILES and chemical text passages in a unified embedding space to support cross-modal retrieval? As described in Section 6.1.5, we focused on the Passage–SMILES modality pair, jointly training two transformer-based encoders (ChemBERTa for SMILES and LongFormer for passages) in a semi-supervised manner using an InfoNCE-based contrastive objective. Hard negative mining is employed to improve discriminative alignment, to pull frequently occurring Text-SMILES pairs away from each other as described in Section 6.1.4.

RQ3 How can the dense encoder learn meaningful representations when a significant fraction of training passages cannot be paired with a ground-truth SMILES due to absent chemical entity mentions, and can pseudo-positive supervision derived from hard negative mining recover or improve retrieval quality under this incomplete pairing regime? While our preliminary formulation addressed missing molecular

diagrams in SMILES–passage using keyword-only search that avoided any contextual training, the current work reformulates this problem for the SMILES–passage setting. Rather than applying a negative-only objective, we re-purpose the hardest in-batch negative SMILES as a pseudo-positive supervision signal for unpaired passages, weighted by a scalar α and combined with the standard InfoNCE loss over valid pairs. This reformulation allows the model to incorporate contextual chemical passages that lack explicit molecular entity mentions without discarding them from training entirely. This process is described in Section 6.1

1.3 Contributions

From the proposed research questions and the approaches to answering them described above, the major contributions of the thesis were the following, grouped into the broad categories below.

Chemical Information Extraction and Indexing

1. A complete end-to-end pipeline for extracting, linking, and indexing chemical passages, molecular diagrams, and SMILES representations from patent PDFs at page-level granularity with full source provenance.
2. A greater than 100-fold increase in indexed passage–SMILES pairs over preliminary work through systematic pipeline improvements.

Graded Relevance Test Collection

1. The first publicly available graded relevance test collection for chemical information retrieval over scientific document data, constructed with domain expert chemists.
2. A novel relevance framework adapted for document-page retrieval that accounts for the distributed nature of chemical information across neighbouring pages.

Semi-Supervised Contrastive Learning with Pseudo-Positive Supervision

1. A hybrid contrastive training objective that leverages both valid passage–SMILES pairs and

ungrounded chemical passages through hard negative pseudo-positive alignment, without requiring ground-truth SMILES supervision for all training samples.

2. Empirical evidence that pseudo-positive supervision can match or exceed a fully supervised baseline trained on twice the paired data volume.

3. Systematic characterization of how loss weighting, temperature, and missing sample ratio interact to govern the benefit of pseudo-positive supervision.

Multi-Modal Query Interface

1. Statistically significant evidence that multi-modal Text+SMILES queries substantially outperform text-only queries, motivating structure-based query interfaces for chemical patent retrieval as a more informative alternative to existing text-only tools.

1.4 Why Can't We Directly Use LLMs or RAG for This Problem?

A natural question that arises in the context of modern large language models is whether the retrieval and indexing infrastructure developed in this thesis could simply be replaced by feeding entire patent PDFs into a sufficiently capable LLM. The answer is no, and for reasons that are fundamental rather than merely practical.

LLMs are generative, not retrieval systems. When a chemist queries an LLM over a document collection, the model produces a generated answer synthesised from its context window. It does not return a ranked list of candidate passages with provenance metadata specifying which page, which location, and which source document each result originated from. This distinction is not incidental — for scientific chemical retrieval, provenance is the result. A chemist who needs to verify a synthesis procedure, confirm experimental conditions, or examine the original molecular diagram as drawn in the patent requires a direct pointer back to the source page, not a paraphrase of its content. Our system returns exactly this: every retrieved candidate carries its source PDF identifier, page number, and spatial location, enabling direct navigation to the original document context. An LLM-generated answer without this page and passage level attribution is scientifically unverifiable or too tedious and therefore of limited utility for expert chemical practice.

Scale and context window limitations. Even the largest context window LLMs available today cannot simultaneously process the tens of thousands of patent documents that constitute a meaningful chemical retrieval corpus. Our indexed collection alone spans 131 patents and over 30,000 pages, yielding 373K passage–SMILES pairs. No LLM can hold this at query time. Processing one document at a time defeats the purpose of corpus-level retrieval, and even within a single long patent, page-level granularity — the ability to identify which specific page of a 200-page document is most relevant to a given query — is not recoverable from LLM generation. Our system performs retrieval across the full indexed corpus in a single forward pass of the query encoder, returning ranked page-level candidates from anywhere in the collection with sub-second latency.

Structure-based querying is not available in LLMs. LLMs with document vision capabilities can read text and molecular diagrams, but they cannot accept a SMILES string as a query and retrieve structurally related passages from a corpus. Our system explicitly supports SMILES-to-passage retrieval, passage-to-SMILES retrieval, and multi-modal Text+SMILES fusion queries, all of which are inaccessible to PDF-ingesting LLMs. As our results demonstrate, SMILES querying is the dominant driver of retrieval quality, producing gains of over 0.38 on R'@10 relative to text-only querying. This structural query modality is the primary interface through which a chemist's expert knowledge of molecular structure can be leveraged at retrieval time, and its absence from LLM-based pipelines represents a fundamental capability gap for chemical information retrieval.

Can retrieval-augmented generation address this? Retrieval-augmented generation (RAG) closes part of the gap by separating the retrieval step from the generation step, but standard RAG pipelines remain inadequate for this problem in several important ways. First, RAG systems typically chunk documents into fixed-size text windows without awareness of page boundaries, spatial layout, or the relationship between text passages and molecular diagrams on the same page. Our system preserves these structural relationships by design, operating at semantically meaningful page-level granularity with explicit passage-to-diagram links encoded through shared SMILES representations. Second, RAG systems using general-purpose text embedders have no chemical domain awareness: the embedding spaces they produce cannot bridge the modality gap between natural language chemical descriptions and SMILES molecular representations, which is precisely what the LongFormer–ChemBERTa contrastive training in this work achieves. Third, and most importantly, RAG has no native support for structure-based queries: a chemist who wishes to retrieve all passages describing reactions involving a specific molecular scaffold cannot express that query in natural language with sufficient precision, and a standard RAG system provides no

mechanism for SMILES-based candidate retrieval or cross-modal fusion.

LLMs as a complementary post-retrieval Re-Ranker: The one context in which LLMs genuinely add value to the pipeline developed in this thesis is as a post-retrieval synthesis layer. Once our system has retrieved the top-k most relevant passages with full page-level provenance, an LLM can summarise, compare, and reason over the retrieved evidence in ways that a pure retrieval system cannot. This is in fact the architecture toward which production chemical informatics systems are converging: structure-aware retrieval followed by LLM-based answer synthesis from the retrieved evidence. This thesis addresses the retrieval half of that pipeline — the harder and less solved problem — providing the page-level granularity, cross-modal alignment, and structure-based query support that LLMs and RAG systems currently cannot deliver on their own.

1.5 Dissertation Outline

This document proceeds as follows. Chapter 2 discusses the detailed background towards understanding the existing work already completed as foundational work as well as set the stage for understanding the scope of the preliminary work. It will discuss the prior work in the chemical information extraction and retrieval domain and introduce the current work including modeling approaches and the training loss functions used for joint multi-modal training. Chapter 3 introduces the currently developed methods of extracting and indexing chemically relevant passages, their associated diagrams and related metadata for enabling in-context PDF search. Chapter 4 dives deeper into our baseline for uni-modal and multi-modal search using term and graph based retrieval, our generated multi-modal test collection containing text and SMILES queries as topics that cater to specific PDFs.

These Chapters lay the foundation for the major contributions in the thesis described in Chapters 5 and 6. Chapter 5 introduces the systems and tools used for increasing the raw extracted passages and molecules and consequently how this helped increase the overall test collection size. This is followed by the collaborative work done with chemical experts to create the relevance graded test collection according to a novel relevance criteria. Chapter 6 delves into the modeling aspect of the work that uses the created test collection to evaluate the new pseudo-positive loss formulation followed by a series of experiments to empirically evaluate the said approach on adding text passages into the training regimen when they do not contain any valid SMILES pairings in text.

Chapter 2

State of Chemical Information Extraction & Retrieval

Chemical Information Extraction (CIE) from documents is the process of extracting and representing molecules, molecule properties, reactions and synthesis pathways in a structured and machine readable way. Extraction can be from passages, figures, tables or charts. In this thesis, we focus on extraction from passages and figures. CIE has always been a challenging domain, both in terms of text entity mentions and molecule diagram parsing. Molecules can be found in an independent context or paired in chain along with other molecules to define reactions. A bigger chain of contiguous reactions can form a synthesis pathway. The challenges arise from the vastly different ways of representing molecules and reactions and reaction conditions in text and diagrams. In addition, available patents and publications in chemistry research before mid 2000's are available solely in a scanned format which means the text is not readily available and needs an extra layer of optical character recognition (OCR). In this Chapter, we attempt to bridge the understanding gap between the text and diagram domains in CIE by discussing representations of chemical name entities in text and molecule diagrams and the methods that exist to represent them in a common SMILES [147] representation.

2.1 Chemical Named Entity Recognition

Chemical Named Entity Recognition (CNER) is a natural language processing (NLP) task of identifying and extracting names of chemical substances from structured and unstructured text such as

scientific papers, patents, clinical reports and web articles. Chemical substances can be found in many different representations such as IUPAC, molecular formula, abbreviation, common and industrial names. CheNER [141] and the system by Campos et. al [12] were one of the first resources in CNER that avoided the use of a complete rule based grammar system and used a Conditional Random Field (CRF) to train a CNER system. However, the model expected full extracted text and only recognized IUPAC [132] names. ChemDataExtractor [138] was designed to be an all-in-one toolkit that also had a CNER component which used a parse-tree grammar rules to detect IUPAC name mentions in text. However, this system too suffered from coverage as it expected text in the form of HTML, XML or born-digital PDFs. ChemDataExtractor2.0 [92] however upgraded CNER to a BERT based model which can now identify IUPAC as well as common names (e.g. prednisone, glucose etc.). OpenChemIE [38] is an end-to-end system that used MolScribe to automatically extract a list of reactions in PDF documents from text and figures, including substituting R-groups (placeholders for atom groups) from tables into corresponding molecules. One important distinction between our method and OpenChemIE in terms of just information extraction is that the latter only used entity linking to substitute R-groups into molecule diagrams. There was no entity linking between text to molecules.

2.2 Molecule Diagram Extraction and Parsing

As a brief overview, most Chemical Structure Recognition (CSR) methods rely on the CLEF-IP2012 [111] dataset for training models to first identify molecule regions in page images. ScanSSD-XYc [34] used a modified Single Shot Detector (SSD) to speed up the process of molecule detection which was subsequently used in math [4] and chemical formula detection [129]. There have been several attempts at parsing standalone molecule figures into a machine-readable representation. The latest work is ChemScraper [129] where they used two different approaches to parsing born-digital and scanned documents to parse chemical structures into SMILES. MolGrapher [99] used a CNN to first predict the atoms and then used a GNN to predict the graph structure of the molecule. MolScribe [114] used a transformer encoder-decoder architecture to directly parse a molecule image into its constituent graph representation before converting the graph into SMILES. Our work improves upon MolScribe to parse chemical diagrams from patent documents. This diagram parser is discussed further in Section 3.2. The following sections delve into some of the key works that have emerged for the domain of molecule diagram parsing.

2.2.1 Rule-Based Parsers

The earliest parser for chemical diagrams in printed documents we know of is a rule-based parser by Ray et. al [121] from the late 1950's. This approach first detected atoms in scanned document images and then connections between atoms were identified in the regions between atoms. Rules based on the number of connections for atoms were used to determine the type of bonds, which worked well for common compounds.

An important later developmentwas the creation of Kekulé system [93]. Kekulé adds additional pre-processing and improved visual detection of bond types over previous methods. Kekulé used thinning and vectorization of raster scans to eliminate variations in bond lines and characters, and ensured that a consistent set of characters and lines were recovered. Once a connection between a pair of atoms was established, the system visually detected the bond type instead of using chemical rules as Ray et al. did. In the same period, CLiDE [49] added the use of connected component analysis in disconnected bond groups to identify bond types. The final adjacency matrix for structure was created similar to Kekulé. Another system by Comelli et al. [20] used additional processing as superscripts or subscripts attached to atoms.

A still-popular open-source system extending the rules of CLiDE and Kekulé is OSRA by Filipov et al. [39]. OSRA refined processing of raster images generated from born-digital documents, which tend to have clearly rendered text lines, characters, and graphics. A similar system is MolRec [127], which uses horizontal and vertical grouping to detect connected atoms, their charge, and stereochemical information. The more recent CSR system [9] also uses rule-based graphical processing to output SMILES representations for molecules, using the *OpenBabel* [106] toolkit to generate a valid connectivity table.

2.2.2 Neural Networks

String Output. Recent advances in neural networks have proven effective for parsing chemical diagrams. For example, Staker et al. [136] use an end-to-end model for extracting molecular diagrams from documents and converting them into SMILES strings. For diagram extraction, they used a U-Net [126] to segment diagrams, which were then passed through an attention-based encoder network [143] to generate a SMILES string representing molecular structure from the segmented image.

DECIMER [118] also uses an encoder-decoder model for extracting molecular structure from raster

images. In their work they explored using different structure representations, including SMILES, DeepSMILES, and SELFIES. They found that SELFIES produced stronger results because of the additional information encoded in comparison with SMILES strings. Additional encoder-decoder parsers include IMG2SMI by Campos et al. [11] which uses a Resnet-101 [45] backbone to extract image features. Li et al. [73] modified a TNT vision transformer encoder [43] by adding an additional decoder. This use of a vision transformer was made possible by the BMS (Bristol–Myers–Squibb) dataset [53] released by Kaggle, which provided a larger baseline for the conversion of molecule images to InChI (International Chemical Identifier names). The training dataset used by Li et al. contained 4 million molecule images. Similarly, SwinOCSR by Xu et al. [155] used the Swin transformer to encode image features and another transformer-based decoder to generate DeepSMILES, and used a focal loss to address the token imbalance problem in text representations of molecular diagrams.

Graph Output. String representations of molecular structure lack direct geometric representation between input objects (e.g., atoms and bonds) and the output strings, and models trained upon them require extensive training data [100]. In recent years, molecular diagram parsers that combine rule-based and neural-based approaches and generate graph representations have emerged. These methods usually employ a graph decoder or graph construction algorithm.

MolScribe [115] uses a SWIN transformer to encode molecular images and a graph decoder consisting of a 6-layer transformer to jointly predict atoms, bonds, and layouts, yielding a 2D molecular graph structure. They also incorporate rule-based constraints for chirality (i.e., 3D topology) and algorithms to expand abbreviations.

MolGrapher [100] is another method employing a graph-based output representation. It utilizes a ResNet-18 backbone to locate atoms, and constructs a supergraph incorporating all feasible atoms and bonds as nodes, which is then constrained. Subsequently, a Graph Neural Network (GNN) is applied to the supergraph, accompanied by external Optical Character Recognition (OCR) for node classification. Both these systems utilize multiple data augmentation strategies, including diverse rendering parameters, such as font, bond width, bond length, and random transformations of atom groups, bonds, abbreviations, and R-groups (i.e., abbreviations for 'rest of molecule') to bolster model robustness.

Likewise, Yoo et al. [157] and OCMR [145] produce graph-based outputs directly from molecular images. Yoo et al. [157] leverage a ResNet-34 backbone, followed by a Transformer encoder equipped with auxiliary atom number and label classifiers. A transformer graph decoder with self-attention mechanisms is used for bonds. In contrast, Wang et al. [145] employ multiple neural

network models for different parsing steps. These steps include key-point detection, character detection, abbreviation recognition, atomic group reconstruction, atom and bond prediction. A graph construction algorithm is subsequently applied to the outputs.

These graph-based methods offer improved interpretability and robustness, and represent chemical structures naturally. In particular, atom-level alignment with input images facilitates easy examination, geometric reasoning, and correction of predicted results.

2.3 Cross-Modal Entity Linking

Cross-Modal Entity Linking is linking one modality of data to another (Image to Text, Text to Audio, Text to Video, etc. and vice-versa). In the context of general document based cross-modal entity linking, the two most prominent works are by Kong et. al [65] and Kim et. al [58]. While the former work generated text-based references to charts in articles and academic papers manually using crowd-sourcing, the latter work designed an automated system for linking text excerpts from publications. Their pipeline first extracted table rows and columns supported by identifying HTML table tags. Then sentences were linked to the table rows based on a hierarchy of rules on set of probabilistic methods – sparse and dense methods to match at the syntactic and semantic levels. The major limitation of their system is that it was designed based on the premise that all research papers have their associated HTML versions as well. This is not the case, specially for the chemical domain where papers can be scanned or are published as patents in which case most patent organizations like USPTO and WIPO do not provide the HTML version. Furthermore, drawn molecule diagrams do not have their HTML equivalent and are usually embedded as images in the PDF.

While there have been no publicly available research in entity linking of textual passages in chemical patents and papers to drawn compound diagrams, the closest work is by Southan et. al [135]. They used the CNER system developed by ChemAxon to identify IUPAC and common chemical compound names from text abstracts and co-referenced them to the PubChem [61] database so chemists could directly get recorded properties of that specific molecule in PubChem through the direct URL linking it.

2.4 Dense Multi-Modal Retrieval

Multi-Modal Learning jointly from text, image and other modalities has evolved over the years. The fusion process of different modalities, objective functions and the size, type of datasets used. In a retrieval context, multi-modal joint training was used for cross-modal setting like MoleculeSTM [78]. However, most prior retrieval works in the chemical domain use a single encoder for cross-modal retrieval like KV-PLM [161], ORMA [96]. In this section, an overview of the early methods and the techniques used will be discussed, from the traditional methods to more neural approaches and finally introducing the transformer based approaches. For the neural and transformer based approaches, the progression of the loss functions will also be discussed as well.

2.4.1 Traditional Multi-Modal Modeling

The earliest work on jointly learning representations from multiple modalities was CCA and KCCA [44]. CCA or Canonical Correlation Analysis developed a linear method where it finds linear projections of two input spaces (Text,Text; Image,Image; Text,Image) $X \in \mathbb{R}^{n \times d_1}$, $Y \in \mathbb{R}^{n \times d_2}$ such that projections are maximally correlated. n is the number of eigen vectors for each modality and d_1, d_2 are the dimensions of each of the eigen vectors for the text and image. The objective function of how this is done and measured will be discussed in Section 2.4.4. It was used for multi-view learning, dimensionality reduction and cross-modal retrieval from text to image and vice-versa. However, the major limitation with this approach was that it could not model nonlinear relationships between two pairs of inputs. To mitigate this problem, KCCA or Kernel Canonical Correlation Analysis was developed, making a nonlinear generalization of CCA. Similar to the kernel trick in Support Vector Machines (SVM), it maps X and Y into a higer dimensional space using kernels $k_X(x_i, x_j)$, $k_Y(y_i, y_j)$. CCA is then applied in that kernel induced feature space. However, the limitations included large computations requiring solving for eigen values on $n \times n$ kernel matrices. Furthermore, KCCA was prone to overfitting without clever regularization techniques. However, these early models showed promise in learning semantic representations between textual captions and images. For instance, a text query "at pheonix sky harbor on july 6 1997" returned images of airplanes in an airport on a hot sunny day. Rasiwasia el al. [120] realized that the image and text shared latent space only has weak semantic relationships in CCA. This is because the text description of an image can contain a lot of words that do not describe the image. They introduced the first concept "Semantic Matching" by combining CCA eigen value decomposition by probabilistic latent semantic analysis by computing the posterior probability through multi-class logistic regression depending on the class of the image and its corresponding description.

2.4.2 Neural Multi-Modal Modeling

The first neural modeling approach on image-text data was done by DeViSE [40]. For encoding text, they used the skip-gram modeling idea on words [95] on a neural model. For images, they used another separate CNN to produce the feature embeddings. The embeddings from the two models were trained jointly using a combination of cosine similarity and hinge rank loss first implemented in [149]. Kiros et. al [63] used VGG [131] to extract 4096-dimensional feature vector from images and parallely used a Recurrent Neural Network (RNN) to encode word sequences. Both modalities were trained jointly using the same hinge ranking loss. Furthermore, they also used an RNN-based decoder to extend the model to generate captions for a given caption as well. MMSKIP-GRAM [67] first used a shared latent space representation between text and images to perform zero-shot retrieval given a text description that contained a concept that was never seen during training. Karpathy et al. [54], was the first to develop the concept of using a unified model for text to image and image to text retrieval using a bi-directional LSTM to encode and generate joint representations of text and images. Section 2.4.4 will delve deeper into the loss functions that have been used throughout the domain self-supervised joint multi-modal training.

VSE++ [37] was the first multi-modal system to introduce hard negative sampling in a multi-modal retrieval setting. They incorporated some of the ideas in the modern day contrastive loss by updating the hinge rank loss from a sum of hinges technique to the max of hinges technique where only the hardest negative example loss was taken in the final loss objective and used a triplet ranking loss. SCAN [68] introduced cross-attention between the modalities to better inform the shared latent space representations. Cross-attention introduces the fusion of features from the two different backbones for the different modalities so that there is interaction between the two featurizers before the output.

2.4.3 Transformer-Based Multi-Modal Modeling

UNITER [17] was the first transformer based multi-modal approach to cross-modal image-text retrieval. They used a Faster R-CNN [122] to first extract features from images and BERT [31] to featurize text. These embeddings were then fed into a common transformer model with both image and text tokens fed into the same unified model. The input and output sizes however were fixed for each modality. This made each output token to have a semantic fused representation of both the image and text context corresponding to its input token. Image and text tokens were separated in the input by a special `[SEP]` token. Furthermore, they showed that the same model can be

fine-tuned for various downstream tasks apart from retrieval such as Visual Question Answering (VQA), Visual Commonsense Reasoning (VCR), Natural Language for Visual Reasoning (NLVR) etc.

CLIP [117], improved upon UNITER by demonstrating that a good alignment can be learnt between image and text modalities without the numerous pre-training alignment tasks that UNITER performed which are discussed in Section 2.4.4. They used a single self-supervised loss called InfoNCE loss as seen in equation 2.6 to learn joint representations between image and text and outperformed all other systems at the time for various downstream tasks.

Compared to UNITER and CLIP which used separate encoders to first encode image and text separately, ViLT [62] showed that just directly using word tokens and patch tokens for images in a unified transformer would perform equally well in downstream tasks, including retrieval. ALIGN [51] showed that increasing the scale of training to a very large pre-training data can help with weakly annotated positive paired image-text data. They empirically show that the effect of weak positive samples can be mitigated by a sufficiently large training set where it is highly likely that most of the positive pairs will be strong. Their approach was to setup rules-based heuristics to filter weak positive paired data. Firstly, they removed all training data where the text modality was too long or short. Secondly, they checked for duplicated data in either of the modalities through similarity matching and filtered them out. This light cleaning procedure ensured that only extremely low quality pairs were filtered out from the training data without losing a majority of the web scraped data. VLMO [6] introduced modality-specific experts to the modeling paradigm. These mixture-of-modality experts (MoME) were added in each transformer block and were basically fully-connected layers separate for each modality in a shared transformer that processed both modalities. This enabled the model (with a single unified transformer) to retain modality specific features while still allowing shared attention. A special router in each transformer block decides which tokens should go to which feed-forward network to better inform inter and intra modal learning.

ComqueryFormer [154] introduced an additional loss to learn local alignment between an image and text. The local alignment was done by selecting special modifier words from the text and tried to align with specific regions in the image. Thereby, the objective was to align a phrase “on top” out of the full text “A dog is sitting on top of a bench” to align well to the region of the image where there is the dog and the bench. Empirically, they found that this combination of global and local alignment allows more semantically rich retrieval across different datasets.

2.4.4 Evolution of Loss Functions for Multi-Modal Training

Socher et al. [133] demonstrated that given training images $x^{(i)} \in X_y$ belonging to class Y can be mapped to a word vector w_y corresponding to the class name through minimizing the L2 distance between the two in a shared embedding space. Formally, given the trainable weight matrix θ, the following objective was minimized:

$$J(\theta) = \sum_{y \in Y} \sum_{x^{(i)} \in X_y} ||w_y - \theta x^{(i)}||^2 \tag{2.1}$$

However, the training objective had a few limitations. It only considered positive examples between the same class of images and texts without actually aligning with other classes. Secondly, L2 distance focuses on just the magnitude between two vectors (vector norm). This is counter-intuitive during training because if the magnitude of the vectors grow too large, the L2 distance will increase even though angle between them is small. Moreover, gradients are prone to explosion if the difference in vector norm between the two modalities becomes too large. This might cause semantically aligned embeddings to be far apart in the latent space because one of them is larger. Later works such as DeViSE [40], Kiros et al. [63] and MMSKIP-GRAM [67], used cosine similarity based hinge ranking loss as shown below.

$$\mathcal{L}(Image, Label) = \sum_{j \neq Label} max[0, margin - \hat{t}_{Label} M \hat{v}(Image) + \hat{t}_j M \hat{v}(Image)] \tag{2.2}$$

$\hat{v}(Image)$ is a column vector for the image embedding, $\hat{t}_{Label}$ is the row vector for the true label and $\hat{t}_j$ is a row vector for negative labels. M is the learned weight matrix. The objective function $\mathcal{L}(Image, Label)$ was aimed to be minimized during the training process. The $max()$ function implements the hinge feature of the loss where if the total loss from the second term is negative, that means that the model is identifying and discerning between the positive and negative pairs efficiently. The positive pair $\hat{t}_{Label} M \hat{v}(Image)$ part of the loss has a negative sign because we want to minimize the total loss if the pair is well aligned to each each (i.e. dot product is large). Conversely, the negative pair term $\hat{t}_j M \hat{v}(Image)$ adds to the loss as their dot product should be as low as possible to signify they are far apart from each other in the latent space ideally. $margin$ is the decision boundary threshold where the difference in loss between the positive and the negative pair has to be at least greater than it. In practice, it was set as 0.1.

VSE++ [37] extended pairwise Hinge Ranking Loss to Triplet Loss with a hard negative mining

strategy as shown below.

$$\mathcal{L}(I, L) = \sum_{I,L} [max_{L' \neq L}(0, \alpha - s(I, L) + s(I, L^{'})) \; + \; max_{I' \neq I}(0, \alpha - s(I, L) + s(I^{'}, L))] \tag{2.3}$$

I and L are positive image and label pairs. $I^{'}$ and $L^{'}$ are the corresponding negative examples. For each positive pair (I, L), the loss finds the hardest negatives $I^{'}$ and $L^{'}$. $s(I, L)$ is the dot product similarity between a pair and α is the margin. If the positive pair is not more similar to each other than the hardest negative by a margin of at least α (usually set to 0.1), the total loss increases. The "triplet" concept comes from the fact that it checks for a positive pair (I, L), both the image and its corresponding hard label $(I, L^{'})$ as well the label and its corresponding hard image $(I^{'}, L)$. This ensured during training that the positive pairs are closer to each other in latent space and the negative pairs are further apart. MSViT [72], adapted the traditional triplet loss to incorporate hard negative sampling in a batch fashion where instead of taking only negative pair in the loss, all the loss generated from a batch of N negative pairs are taken in the loss.

With the adoption of transformer-based multi-modal models, different optimization functions evolved. Without explicitly training a model for a single task, models were pre-trained on large scale datasets in a self-supervised manner first, and then fine-tuned on a smaller dataset in a supervised manner for a range of downstream Vision + Language (V+L) tasks such as Visual Question Answering (VQA) [139, 83, 71], Visual Commonsense Reasoning (VCR) [159], Natural Language for Visual Reasoning (NLVR) [137], Multi-Modal Natural Language Inference Visual Entailment (SNLI-VE) [88] and Image-Text Retrieval [112, 77]. The pre-training tasks were the main tasks to align the image and text modalities together.

The main pre-training task is Masked Language Modeling (MLM) in a multi-modal setting with the objective function described below.

$$\mathcal{L}^{MLM}(\theta) = -\mathbb{E}_{w,v \sim D} log P_{\theta}(w_m | w_{m'}, v) \tag{2.4}$$

θ are the model parameters being tuned, $(w, v) \sim D$ signifies the text tokens w and visual tokens v sampled from the dataset D. w_m represents the masked words in a sentence and $w_{m'}$ represents the unmasked words. $P_{\theta}(w_m | w_{m'}, v)$ represents the predicted probability of a masked word w_m token being a specific word given the unmasked word $w_{m'}$ and the visual token v. $\mathbb{E}_{w,v \sim D}$ represents the expectation over the dataset distribution D, when a random word and a visual token is sampled.

The objective function trains the model to infer masked words given both the linguistic context and visual features. It encourages the model to jointly learn context from textual and visual information.

Another joint training task is called Image-Text Matching (ITM). In this task, the additional special token `[CLS]` is fed into the model which unlike uni-modal models now contains semantic context from both modalities. The inputs to the model is a set of image regions as visual embeddings and sentence. The task is to perform a binary classification – for each image region, whether the sentence describes the patch. Mathematically, it is a binary cross entropy loss (BCE) which is minimized.

$$\mathcal{L}^{ITM}(\theta) = -\mathbb{E}_{w,v\sim D}[y\, log\, s_\theta(w,v) + (1-y)\, log(1-s_\theta(w,v))] \tag{2.5}$$

(w, v) are word and visual token pairs and depending on the ground-truth label y, either the first term is 0 (when $y = 0$) or the second term is 0 (when $y = 1$).

There are other pre-training tasks in the transformer-based multi-modal space like Word-Region Alignment (WRA) and Masked Region Modeling (MRM) but the ones discussed above are the major ones predominantly used.

Moving away from having separate pre-training tasks for learning joint representations between modalities as a self-supervised approach, Oord et al. modified the max hinge ranking loss and triplet loss commonly used before to InfoNCE loss [142]. InfoNCE loss stands for Information to Noise Cross Entropy loss and can be mathematically defined as follows:

$$\mathcal{L}^{INCE}_{i,t} = -\, log \frac{exp(s(i,t^+)/\tau)}{exp(s(i,t^+)/\tau) + \sum_{j=1}^{N-1} exp(s(i,t_j)/\tau)} \tag{2.6}$$

$s(i, t^+)$ is the computed similarity between an image and its corresponding positive text. $s(i, t_j)$ is similarity between the same image and other texts that are not related to it. These become the negative pairs. N is the mini-batch size, where there is one positive pair and $N - 1$ negative pairs for an image. To promote learning from both modalities, in practice, the InfoNCE loss from both Image to Text ($\mathcal{L}^{INCE}_{i,t}$) and Text to Image ($\mathcal{L}^{INCE}_{t,i}$) are computed and averaged. CLIP [117] was the first work that applied InfoNCE loss in a multi-modal transformer based setting. This setting is still most common loss function used in closely related multi-modal works like BLIP, BLIP-2 [70], ViLT [62], VLMO [6], ReViz [84].

A few works after this extended the InfoNCE loss for three or more modalities. Instead of averaging two losses for two modalities, it was six pair-wise losses that were computed and averaged in a 3-choose-2 way if there were 3 modalities and so on. The only works that have implemented this to the best of our knowledge are MURAL [50], MERLOT [160] and M3AE [5]. However, all of these works only implemented this technique in a general domain setting with images, texts and audio or video.

This work will be the first attempt to apply InfoNCE loss in a chemistry domain specific context and further extend it to account for missing positive examples from a particular modality.

2.5 Chemical Information Retrieval

Retrieving relevant information from scientific documents requires two essential pieces – a reliable domain specific extraction system to index information from text, figures, tables or charts and a retrieval algorithm for search. Sparse probabilistic models like BM25 [124] and TF-IDF have evolved into dense neutral models like ColBERT [57], DPR [55] and ANCE [153] for searching domain specific information in the Math, Biology and Chemistry based documents. For this work, we will focus on general multi-modal retrieval as well as specialized chemical retrieval.

Multi-Modal retrieval has many facets to it. Models like CLIP [117], CLIP-Branches [86], BLIP [70], ALIGN [51], UNITER [17] are dense models that focus on aligning image and text modalities in a latent space. While ALIGN has a dual encoder design and generates separate vector embeddings for image and text, UNITER uses a joint-encoder for both modalities in the same latent space and is better for reasoning tasks. Most of these models provide baselines for the ImageNet [29], MS-COCO captions [16] and Flickr30k [158] datasets they focus on the cross-modal retrieval task: image $\rightarrow$ text, text $\rightarrow$ image.

The retrieval methodology for this work is fundamentally different. This is because unlike these datasets which have fixed text-image pairs as independent samples (suited for exact item retrieval), there are many-to-one and one-to-many correspondences in the test collection. Some passages may be linked to more than one drawn molecule figure. Conversely, some compound diagrams maybe associated with more than one passage. Currently, there do not exist dense retrieval models designed for searching chemical documents. This work is a first attempt at creating a baseline for this challenging task.

The first commonly known chemical search system was ChemXSeer [97] that extracted chemical

names from tables in PDFs and indexed them. It was a simple system that extracted all chemical names from PDF tables and indexed them to be searchable by name or chemical formula. TREC-CHEM [85] was an early research competition that manually curated patents and used chemistry graduate students to create relevance assessments. This competition only looked at documents at the PDF level and was a good first attempt at creating a chemistry retrieval collection. Edwards et al. developed Text2Mol [36] that used information available in the PubChem [60] and ChEBI [28] databases to train perceptron and GNN models on SMILES-Description pairs where the system can recommend the best SMILES given a user description of a molecule. MoleculeSTM [78] was a similar concept but used a multi-modal transformer network to jointly train molecule and description pairs with contrastive loss using a transformer encoder network. However, their retrieval task was not a traditional task, but picks the most likely text description out of N descriptions (multiple choice) given a compound as SMILES.

Later works such as SureChEMBL [107] made an attempt to index chemical compounds from both text (IUPAC Names) and drawn molecules together from chemical patents. But their extraction and indexing process had several limitations that our work addresses.

- They index unique compounds found in text or diagrams as a single instance, losing the multi-modal aspect of the extraction. In contrast, this work indexes all diagrams and chemical name instances separately, and also links passages with their associated diagrams.
- Only text appearing in patent titles and abstracts are indexed, along with a set of unique molecules that may appear as name mentions or in drawings mentioned previously. This work indexes all text passages with chemical name mentions, and all extracted molecule diagrams.
- Extracted IUPAC compound names can only be searched using exact matching. For e.g., searching for *fluorooxetan pyrimidine* will yield no hits even if a molecule named *2,4-dichloro-5-(3-fluorooxetan-3-yl)pyrimidine* exists in the index. This work uses a specialized tokenizer that divides IUPAC names into functional group tokens, supporting bag-of-word matches on IUPAC tokens.
- Retrieval is performed only at the document level, as opposed to the passage-level search within PDF documents that this work supports.

In summary, in comparison with SureChEMBL, UniChemFinder indexes all extracted chemical instances, uses a lower level of granularity to represent compound names to support more flexible matching, and results are returned at the level of passages in PDF documents, rather than a list of links to documents matching the query.

Reaxys, SciFinder [41] and the system developed by Akhondi et al. [3] suffer from many of the same limitations as SureChEMBL. While Reaxys and Scifinder do support hybrid multi-modal querying, both use strict boolean comparisons and ranks hits based on keyword matches. CHEMDNER [66] introduced a corpus of about 80k compounds manually extracted and annotated according to their chemical name entity mention class (IUPAC name, trivial name, etc.) from PubMed abstracts. They required chemical experts and mined only text abstracts. This dataset was used to train many CNER models. Many other corpuses such as DrugBank [150], ChemIDPlus and ChemSpider [110] extract substances from text in patents and publications and provide a cross-reference back to the patents they were originally found in. However, these are not complete solutions and do not provide the level of granularity that chemists need including indexing drawn figures and functional-group level tokenization. The major reason is chemists often look for additional contextual information about their search like pharmacological properties, reaction conditions, etc. that are generally found in and around spatial location of the candidate hit within the PDF page. PatCID [102] used MolGrapher [101] to create an automated system to index drawn molecule diagrams and developed a search system for only drawn molecule diagrams but suffers from same granularity and completeness problem as the other systems.

The major constraint of these systems is that they do not allow searching by functional groups, thereby implicitly losing the ability to search structures by parts. SureChEMBL, Reaxys and SciFinder do allow a user to manually draw structures for searching but they have to be complete structures. For example, one can only search for compounds containing 'pyrimidine' as a sub-group but not 'pyrimidine' independently. Section 4.1 demonstrates how we enable this using a chemistry inspired tokenization scheme for IUPAC names.

2.5.1 Benchmark Datasets Used in Chemical Information Retrieval

There are four main datasets used that are used to benchmark retrieval systems in the chemical or biochemical domain. Most of the created benchmark datasets use the PubChem database to create molecule-description pairs.

- The earliest dataset that was created was the **PCDes** dataset [161]. It contained 15k Canonical SMILES-Description pairs. The molecule property description was mined from PubChem's "Description" and "Use & Safety" section for each molecule metadata. Training, validation and test splits were made randomly into a 70-10-20 ratio. It is worth mentioning that the majority of molecular retrieval benchmark papers reported their retrieval scores on

this dataset.

- **ChEBI-20** [36] was created by first scraping molecule annotations from ChEBI [27] and then cross-referenced with PubChem. Through this process, a subset of 33k molecule-description pairs was created.
- The next work was a more thorough approach using Large Language Models (LLMs) and expert evaluators to construct a set of 300k molecule description pairs [163] which is commonly called **PubChem-300k**. They used GPT using a template of prompts to generate properties from a given SMILES and conversely SMILES from a given set of properties. This initial set was then expertly evaluated and incorrect, duplicate or chemically ambiguous samples were filtered out.
- The most recent work on creating a benchmark dataset is the **MolTextNet** dataset [165]. Compared to PCDes which scraped only the Description and Use & Safety sections from PubChem, MolTextNet scraped a much larger subset of PubChem to generate 2.5 million high quality molecule-description pairs. They used additional metadata sections such as "Chemical and Physical Properties", "Use and Manufacturing", "Pharmacology and Biochemistry" and "Toxicity".

2.6 Commonly Used Search Platforms Today

Although systems like SureChEMBL and PatCID have been important steps towards making a unified search platform for substances and obtaining their molecular property data, their widespread adoption by chemists have been limited due to their scope of molecules and reactions covered as well as a lack of linking of compounds back to their original PDFs. Chemists and biochemists prefer to use a semi-automated curated collection of compounds and reactions found in Reaxys and SciFinder due to its large coverage and association of compounds with the patents or papers they were found in (at an overall level).

Figure 2.1 shows an example search in the SciFinder system. The initial search result for a molecule (*2,4-diamino-5-(3,4,5-trimethoxybenzyl)-Pyrimidine*) within patents shows the title and the text abstract as a hit whether or not the actual searched molecule was mentioned in the abstract or not. Moreover, on navigating to the substances results as shown, it only shows molecules in the PDF sorted by similarity. There are no similarity scores mentioned or the location in the PDF where the compound was indexed from. This becomes a bottleneck for chemists as they are not

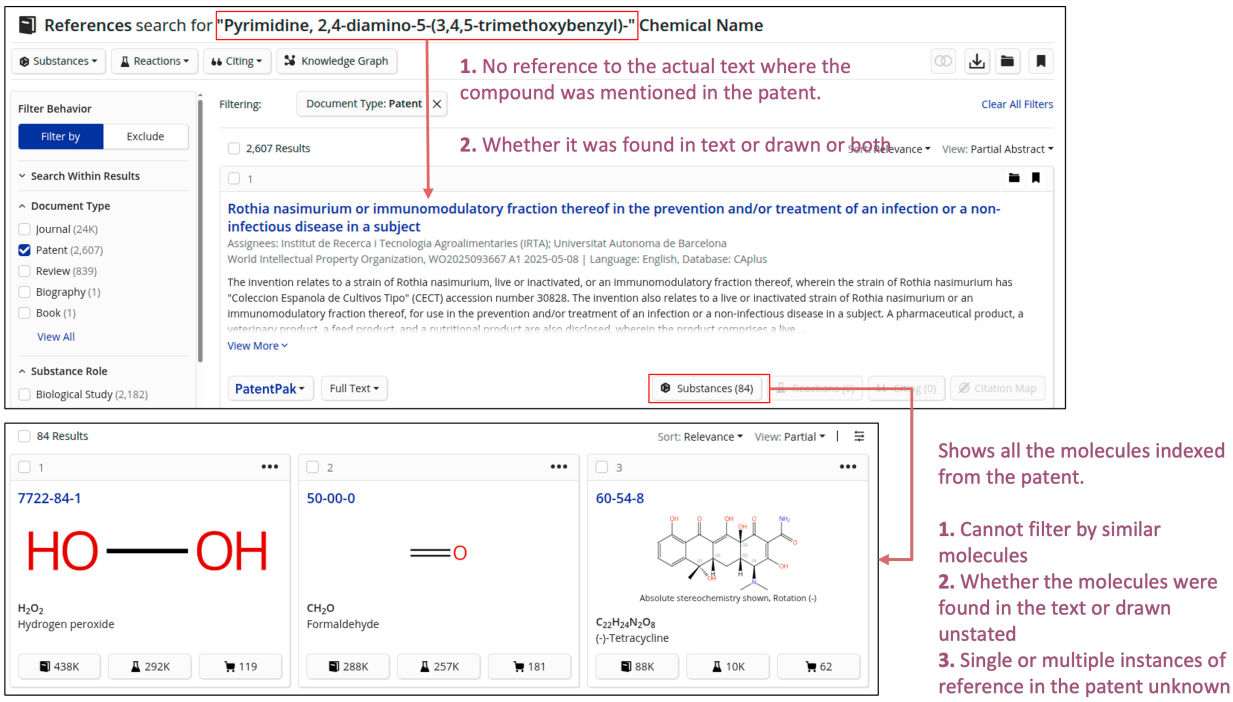

Figure 2.1: Example Search of *2,4-diamino-5-(3,4,5-trimethoxybenzyl)-Pyrimidine* in SciFinder. The left navigation pane provides the users ability to filter by source (journals, patents, etc.), substance role (reactant, product, reagent, etc.) and authors. Overall, the system does a similarity search over all molecules in the index and does not offer advanced filtering methods such as exact matches only, sub-structure match only or provide passage and page level references to where the molecules were referred to in the PDFs

only looking for similar compounds but sometimes the criteria of similarity might be different for different needs. Furthermore, they have to manually search the entire patent/journal physically to look for specific property data that they might be interested in for various downstream tasks. This is partly because SciFinder and Reaxys compiles a common database of property data of molecules and it is not certain whether the specific patent in question was used to compile the data or not. In addition to missing data, there might also be a challenge of stale or old data associated with a molecule. This is because, chemists perform experiments using the same molecule all the time and update any results that they find. This forms the basis of our motivation to mitigate a lot of the limitations of such systems by allowing chemists search patents for compounds by different similarity criterias and showing results at the page and passage level for every PDF that the user can navigate quickly and efficiently to.

Chapter 3

Extraction and Linking of Chemical Passages & Diagrams

We demonstrate through this work that enabling page and passage level search of either text or drawn molecules in literature is a useful addition to the existing Chemical Information Retrieval (CIR) tools like Reaxys and SciFinder. Current CIR systems for literature search from documents have traditionally treated the domain as two distinct sets of sub-problems: molecules as text or images. One involves techniques for Chemical Named Entity Recognition (CNER) from text and the other involves parsing a 2-D molecule diagram into a machine readable representation such as a SMILES string or a MOLFile. Moreover, the underlying assumption was that creating an extracted set of compounds or reactions in isolation as a molecule bank without cross-linking back to the original sources is sufficient. Additionally, for literature search, complete text was indexed for only abstracts. Thereby, searching for very specific compounds using their properties or reactivity constraints is not possible as these descriptions are not likely to be found in abstracts.

This work attempts to mitigate these limitations by creating an extraction system where passages in text containing chemical information are first identified. Drawn molecule compounds are identified and parsed into their SMILES representations. Molecule names and diagrams are linked across a PDF to produce passages with their associated drawn molecules. Overall, searching for relevant chemical information is enabled through indexing only passages with molecule names linked to their associated molecule diagrams.

3.1 Text Extraction

The text extraction pipeline developed in this preliminary work integrates a number of processing systems together. These systems detect text blocks, parse the words in the block into a machine readable form, detect chemically relevant terms in each text block and filter out blocks that are not chemically relevant if they do not contain a molecule name mention. The main motivations for this pipeline are:

1. Many patents and journal publications do not have a corresponding HTML or born-digital version from which passages and text can be readily parsed. Therefore, there was a need to create a chemical text extraction system from PDF collections – whether scanned or born-digital.
2. Commonly used chemical literature search systems such as Reaxys and SciFinder only index full text passages for abstracts. This can limit searches that chemists do for specific reaction conditions along with molecule information. We index all passages from documents that have at least one valid chemical name entity in them by identifying any IUPAC names within it.
3. To avoid manual search in a PDF (for specific searches and finding additional associated information in existing tools), in our work proposes we keep a record of the identifying unique PDF, page and location information for each passage candidate. Therefore searches can directly point to candidate passages within a PDF.

The following Sections below will describe the text extraction pipeline components.

3.1.1 Passage Detection and OCR

The text extraction pipeline comprises of two main steps:

- *Passage Segmentation*: Each PDF was first converted to its corresponding page images and passed through *LayoutParser* [130]. For segmenting passages from images, a Mask-R-CNN model was used that was trained on PubLayNet [164]. PubLayNet consists of journal articles and preprints in the PubMed Central Open Access Subset [103]. After passing all the PDFs through the segmenter, passages in the images were segmented out and their corresponding metadata such as the PDF, page and location recorded.

- *Optical Character Recognition (OCR)*: The individual passage segments were passed through PyTesseract [56] to extract their text content. It was important to make sure that long IUPAC or common chemical names that were broken down into two separate writing lines were joined according to chemical rules. Specifically, if any line began or ended with a "-", they were joined with the hyphen in place. This is because in chemical articles, hyphens are strategically used to break long chemical names into two lines where there is supposed to be the separator, for e.g., `2,4-diamino-5-(3,4,5-trimethoxybenzyl)-Pyrimidine` could be separated as "2,4-diamino-5-(3,4,5-" in one line and "trimethoxybenzyl)-Pyrimidine" in the next line.

3.1.2 Chemical Named Entity Recognition (CNER)

The raw passages extracted as text blocks previously may or may not contain any reference to chemical entities. To keep only chemically relevant passages in the index, we filter passages based on whether there are any IUPAC names that are contained in the passage. ChemDataExtractor2.0's [92] Chemical Named Entity Recognizer was used for this task. ChemDataExtractor2.0 uses a combination of a linear-chain Conditional Random Field (CRF) based model and a dictionary based recognizer [48] to identify IUPAC names, trade and trivial names as well as a regular-expression-based recognizer for chemical formulas. For instance, this CNER system can recognize all forms of common salt in passages like "NaCl", "Sodium Chloride" and "Salt". However, only passages with IUPAC names are indexed in the collection.

While the entities detected by ChemDataExtractor2.0 are comprehensive, we only select passages with IUPAC names. This is because of the motivation to enable searching compounds by its canonical Simplified Molecular Input Line Entry (SMILES) [147]. SMILES is a string based encoding of any molecule from which the original molecule can be constructed back. It not only encodes atoms and bonds between them but also stereochemical properties like chirality, hybridization etc. For eg. the compound benzene (C_6H_6) can be represented in its canonical form as `C1=CC=CC=C1`. This is motivated by chemist needs, information availability and convenience such that any molecule can be searched through a string rather than forming the IUPAC name or drawing out the full structure explicitly. Furthermore, substructures can also be searched through this representation. Although there exists systems to convert IUPAC names to SMILES, common names, trivial names and molecular formulas rely on manual links to be converted. The ambiguity if writing molecular formulas also adds to the complexity. For eg., benzene can written as C_6H_6 as well as H_6C_6. For these reasons, this work currently only indexes passages with valid IUPAC names.

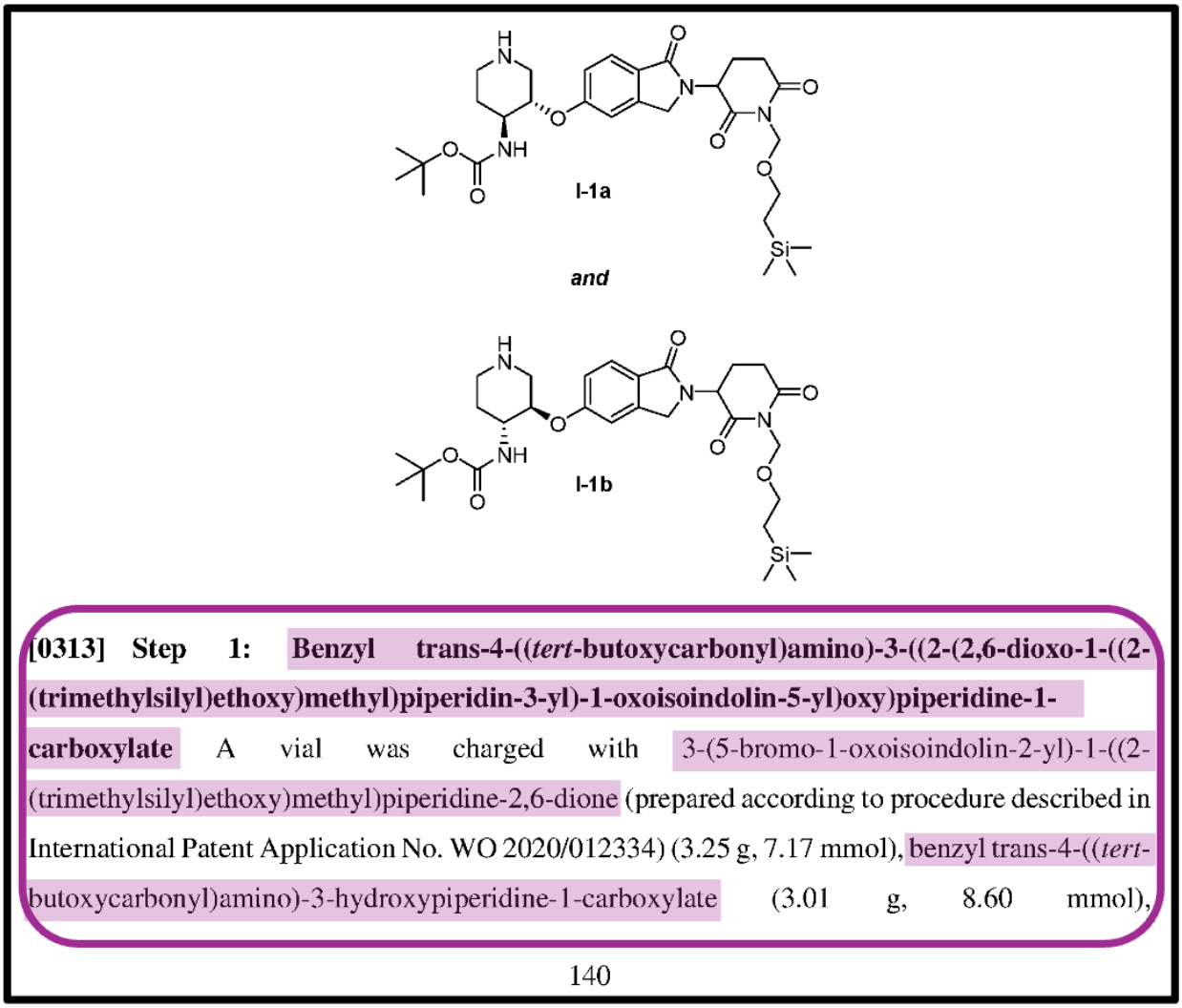

Figure 3.1: A passage in a PDF document with valid chemical names has been successfully detected with LayoutParser (bold maroon bounding box) and the text has been extracted through PyTesseract. Finally, ChemDataExtractor2 was used to identify word sequences that corresponded to a chemical entity (highlighted words in maroon) that form individual compounds.

3.1.3 SMILES Conversion

Once the passages with only IUPAC have been identified and others have been filtered out, the Open Parser for Systematic IUPAC Nomenclature (OPSIN) [81] was used to convert and link each chemical entity in a passage to its constituent canonical SMILES. OPSIN uses a specialized IUPAC tokenization scheme and a hierarchy of rules to construct the encoded SMILES for a IUPAC

name. For instance, the IUPAC name "methyl 1-bromothieno [3,2-f] quinoline-2-carboxylate" is converted to its corresponding SMILES, `BrC1=C(SC=2C1=C1C=CC=NC1=CC2)C(=O)OC`. Section 3.3 demonstrates how these SMILES are utilized to further link passages with their associated figures in the PDF.

3.2 Molecule Diagram Detection and Parsing

The detection and parsing system first identifies the regions where the molecules diagrams exist, then parses the individual molecules through a transformer based encoder-decoder model to predict the graph structure of the molecule. The graph is then post-processed to generate the canonical SMILES string. This representation is useful for a variety of reasons:

1. Applicability in molecular downstream tasks
2. Encoding of chemical properties such as chirality (two similar molecules not superimposable on their mirror images due to differing 3-D bonds)
3. Easy interconversion to other formats such as MOLFile, DeepSMILES [105] and InChI [46, 47]

Molecule Diagram Detection & Parsing

PDF pages are processed through YOLOv8 [52] to segment the molecule regions. CLEF-IP2012 [111] [34] dataset consisting of 1242 pages for training and 419 pages for testing was used to train the segmentation model. Due to limitation in availability of open source training data for segmentation, data intensive options like vision transformers were avoided for this stage.

The parsing model developed, MolScribeV2, improves upon the original MolScribe [114] in three major areas – enhanced positional embeddings, dataset size and special augmentation. Figure 3.2 shows the additional positional embedding method developed that is created from the binary mask of graphics pixels that forces attention values in the transformer layers to distinguish between valid and invalid pixels. The motivation behind this was that the majority of pixels ($> 90\%$) in a molecule diagram image are background pixels and should have no relevance to the feature extraction. The diagram parser takes in segmented molecule images and produces a graph representation of the molecule with the nodes as hidden carbons or explicit characters or sub-atom groups. The bonds are represented as edges in the graph with their types as either: (1) Single (2) Double (3)

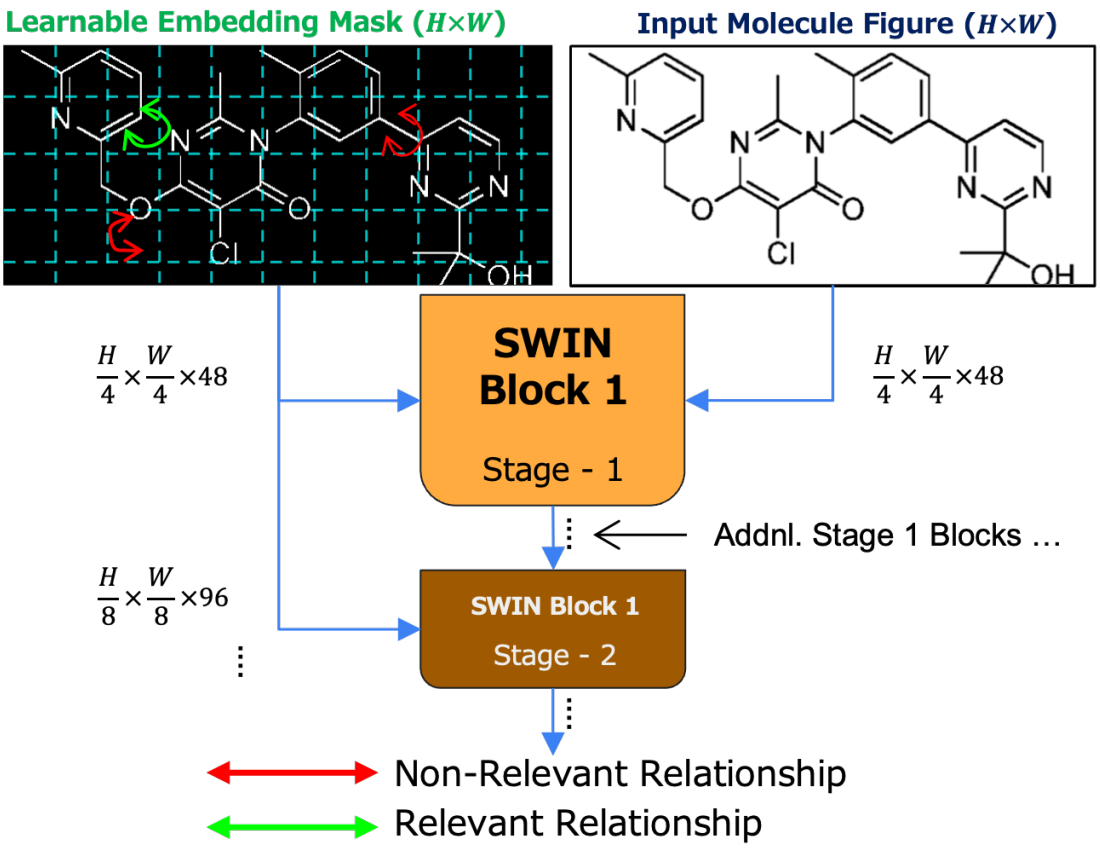

Figure 3.2: Enhanced Learnable Positional Embedding at the SWIN-B encoder for MolScribeV2. A binary mask of the molecule is generated for each of the transformer block levels. For each window of the mask, self-attention layers are generated where pixels belonging to the molecule lines are considered as "relevant" (green arrow) while all other relationships are made "non-relevant" (red arrow). The mask considers pixel relationships within each window.

Triple (4) Chiral. An external dictionary is used to break down sub-group abbreviations into their constituent atoms (E.g.: "Ph" or phenyl group is broken down into its constituent atoms and bonds corresponding to C_6H_5). Finally, this graph is encoded into a MOLFile and converted to its canonical SMILES form using RDKit. A MOL is a molecule encoding format which encodes in a graph representation, each atom as a node with their relative spatial positions and any bonds between the nodes as an edge in the form of an adjacency matrix.

The binary pixel mask is generated by binarizing the original image and identifying the pixels that correspond to the bond lines and characters in the molecule image. An OTSU filter was used to

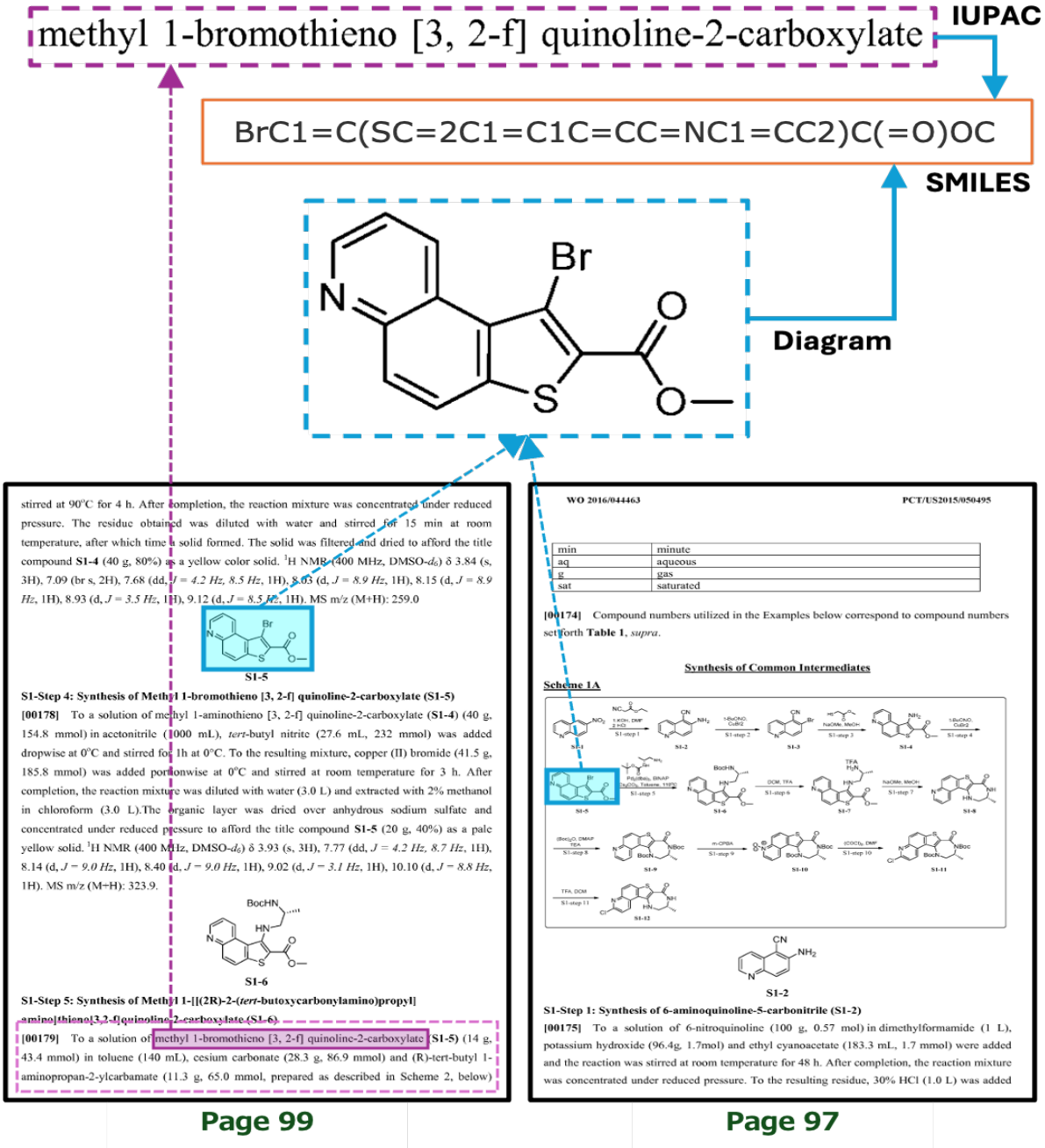

Figure 3.3: Linking the text and diagram modality is done through a common SMILES representation. Maroon represents information extracted from text and cyan represents diagrams. The passage detected on page 99 of a PDF shows the IUPAC name mention '`methyl 1-bromothieno [3,2-f] quinoline-2-carboxylate`' which was successfully converted to SMILES through OPSIN. This name was matched with its corresponding figures at two locations in the PDF (pages 99 and 97) using Tanimoto Similarity on the SMILES generated from the two modalities.

adaptively binarize each image. A SWIN-B [80] transformer encoder was used to extract features from the molecule image. SWIN (Shifted Window Transformer) is a type of vision transformer that computed pixel attention within an image in a localized window which is usually a 7×7 patch. At each each progressive block, the window shifts half the size of the window down and right thereby efficiently computing global attention without the need of full global attention (every pixel to every other pixel) at each block. All the default parameters used in the original SWIN architecture were kept fixed including the window sizes of 7×7 and the number of transformer blocks in each stage $(2, 2, 18, 2)$. The SWIN transformer has a total of four stages each with its own number of transformer blocks. Attention is constrained within each window and the window shifting at each alternate each block ensures attention is not completely local. At each block at each stage of encoder, the positional embedding mask was scaled to the current feature size as shown in Figure 3.2 and added to the relative positional embedding mask before the self attention computation. The quantitative improvements by this change is discussed in Section 4.5.

Compared to MolScribe which was trained on 1 million PubChem [60] molecules, 5 million molecules were used to train MolScribeV2 with data from PubChem, Zinc250k [2], ChemBL [94], MOSES [113]. Finally, existing data augmentation methods were combined with two new techniques to improve performance on real world data. Margin cropping or adding (removing or adding either the top, left, down or right edges to the detected molecule regions) was a method used to cut or append margins (upto 10% of height or width of the image) at any side of the original to simulate real world examples of images arising from imperfect segmentations from YOLOv8. Document degradation was another augmentation that added random amounts gaussian and salt and pepper noise to the images to simulate molecule regions from scanned PDFs.

3.3 Entity Linking

After both passages containing chemical entities and molecule diagrams have been successfully extracted and indexed including PDF location metadata, this step is used to link any passage with all references to drawn molecules that it mentions anywhere in the PDF as shown in Figure 3.3. After the linking process is complete, an additional index is created that contains reverse links from drawn molecules to all the passages they are mentioned in. Appendix B provides a detailed reference to all the created indexes. A passage can contain more than one chemical entity and each of those entities can be linked to more than one drawn figure anywhere in the PDF. Essentially, the linking process is a many-to-one or one-to-many linking where both methods are unambiguous in nature.

Associating passages with their related diagrams in each PDF is done in an unconstrained fashion by taking advantage of a statistical measure of molecular similarity given by something known as Morgan Fingerprints [98, 90]. This removes the usual constraint of entity matching in documents by page or proximity to reduce the number of false positives or false negatives. That is, the most common approaches to this is based on spatial proximity. In our case, it would have been linking drawn molecules to the closest passage to it in terms of distance. In a chemical context, this is important as in many cases, passages can mention molecule diagrams that are drawn in pages far apart.

Fingerprints are essentially a fixed length binary vector that encodes characteristics of a molecule in terms of graph traversals starting from each node and traversing its neighbors. We measure similarity between a pair of fingerprints using Tanimoto Similarity [13]. In the matching process, each passage SMILES represented by molecular graphs is compared to every other diagram SMILES using their graph representation and those having a Tanimoto Sim. greater than 0.95 are recorded as true matches. We do not maintain a hard constraint of 1.0 as there are a few cases where due to improper segmentation of molecule figure regions, the SMILES may have some extra characters which are from a different closely located figure. In practice we have seen the true matches to be precise by using this threshold without adding any false positives.

Chapter 4

Term & Graph Matching Based Chemical Retrieval

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 [132] 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 [97], TREC-CHEM [85], Text2Mol [36], MoleculeSTM [78] and SureCEMmBL [107]. 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.

This chapter starts by describing the inverted indexes that are created to facilitate efficient uni-modal and multi-modal search. These indexes enable searching PDF pages with any mode of search query — text or SMILES or multi-modal. Afterwards, the modes of search are elucidated

with search models for text and SMILES search along with the special way of tokenizing the IUPAC names. The fusion method for merging text and SMILES-based search for the multi-modal queries is also discussed.

Three different models have been developed as an initial baseline system to evaluate the pipeline. The baseline text search uses a sparse BM25 [125] scoring model, while the SMILES based search is done through two different methods provided by RDKit. The multi-modal method involves fusion and re-ranking of the two results obtained from text and SMILES based models. While the potential of dense retrieval models and their perceived performance gains over the currently used sparse models are self-evident, starting with a sparse modeling approach on our new test collection—described in this chapter—provides a strong foundation for this work, as it will empirically demonstrate the performance gains achieved when more contextual representations are used for indexing and retrieval.

4.1 Text Search

BM25 has been chosen as the baseline text search model. Based on the unique passages present in the Passage Index described above, BM25 scores each passage for a given query as a single document. The standard BM25 model is used from PyTerrier [89] with default parameters. Term frequency saturation k_1 is kept at 0.9 and the length normalization parameter b is kept at 0.4. No meaningful impact to the results discussed in Section 4.5 were observed by tuning these parameters.

The important addition to the existing model is the use of a chemically-informed text tokenization scheme. Inspired by STOUT [119] where they tokenized IUPAC names to predict chemical names from molecule SMILES, we adapt their scheme, breaking down IUPAC names into their constituent groups. We further augmented the group tokens by additional ones which are marked by the superscript 1 in the list below. The complete scheme is as follows:

Opening/closing fence symbols: `{, (, [, }, ), ]`

Punctuation: `.  , -`

Common prefixes/intermediates: `mono, di, tri, tetra, penta, hexa, hepta, octa, nona, deca, oxo, methyl, hydroxy, oxy, chloro, cyclo, amino, bromo, hydro, fluoro, methoxy`[1], `ethoxy`[1], `phenoxy`[1], `methane, cyano, amido, ethene, phospho,`

[1]Additional Tokens compared to STOUT

`carbono, hydro, sulfino, ethano, methano`[1], `din`[1], `iodo, ethane, ethyne, bi, tri, iso, nitroso, hydro`

Common suffixes: `benzene, sulfane, methane, ethane, butane, amine`[1]

Given a document text, regular expressions are used to break input text into their constituent chemical text tokens. The parentheses and punctuation marks can be removed in the final string, with little loss of information. Thus, the IUPAC name

```
6-(3-methoxyphenyl)quinazoline-4-amine
```

is tokenized as

```
6 3 methoxy phenyl quinazoline 4 amine.
```

This form of tokenization in a chemically sensitive manner by functional groups within IUPAC names enables searching the passage index by sub-parts of molecules as text compared to a SMILES string. This functionality is useful to chemists, especially in the context of drug discovery where a lot of their information need is to find specific sub-structures (sub-groups) of interest that are known to help inhibit the activity of a certain gene by binding to its protein structure. For example, if a user searches for sub-group "pyrimidine", any full compound in the index that has the sub-group in them (e.g., 1-trifluoromethyl pyrimidine, 2-bromo-5-fluoro-pyrimidine) maybe to be valid candidates for the query. However, in expanding to this tokenization form also makes the search space too ambiguous and there might be passages that are irrelevant in the context of the user.

We discuss in Section 6.1.5 how we will address the issue of relevancy by using dense embeddings in a common vector space that will take context into account when generating hit candidates.

4.2 SMILES or SMARTS Search

We solely rely on RDKit for SMILES-based search. RDKit is a cheminformatics toolkit that can be used for converting molecules from one form to another (e.g. SMILES to MOL, a graph based 3-D representation), determining chemical properties of molecules as well as sparse vector based similarity search. We use RDKit both as a canonicalization tool to make sure that the SMILES

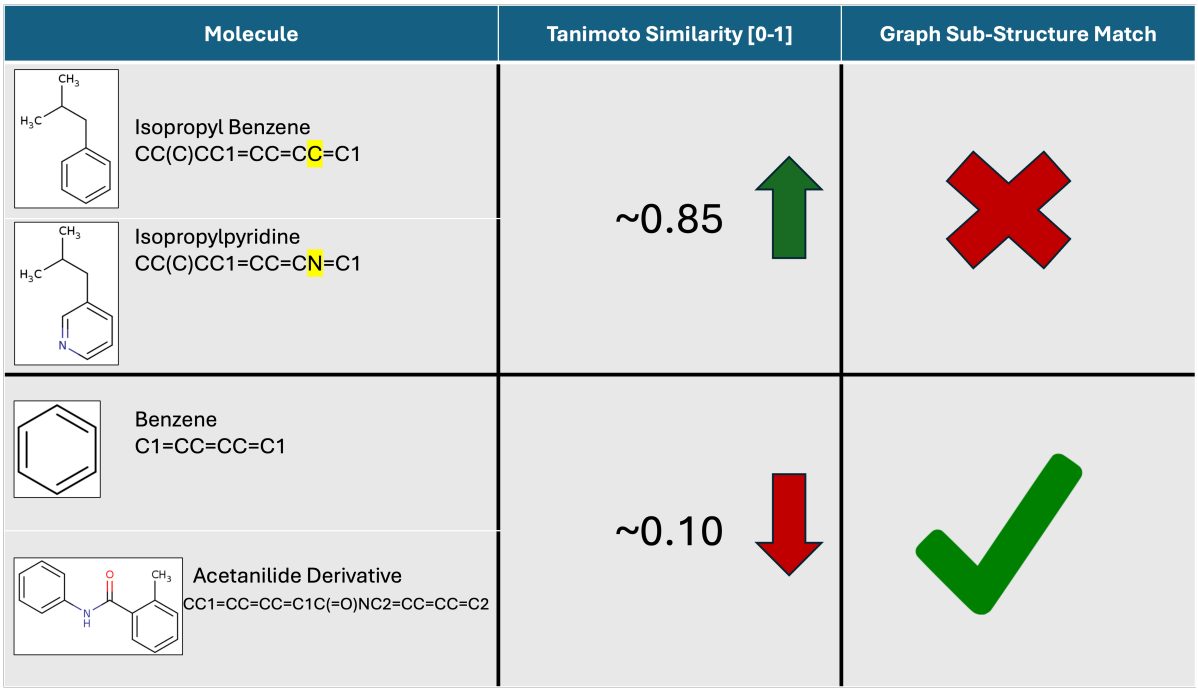

Figure 4.1: The difference between Tanimoto Similarity and Graph Sub-Structure Matching is shown between two pairs of molecules. In the top pair, as only one atom is different between the two, Tanimoto Similarity is very high but the full graphs do not match. Conversely, in the bottom pair, even though the two structures are very different to each other in composition, there exists a common substructure (phenyl ring) between two and sub-struture shows a positive match.

character ordering of compounds from text and diagrams follow the same scheme as well as a standalone retrieval model to search for similar SMILES.

Sub-structure Search Model (Graph-Based)

Sub-structure compound search refers to searching for a sub-group of a compound from an index of full molecules. In our case, the test topics are designed with this in mind as the motivation is to enable chemists to search for relevant molecules through sub-structures of interest, preferably with an accompanying text query to find more relevant results. Sub-structure search in RDKit works as a subgraph isomorphism problem where it is only checked whether there exists a mapping between the query atoms and bonds to a subset of nodes and edges in the target graph.

Tanimoto Similarity Search Model (Vector Based)

Molecule SMILES are converted into their binary vector representation through Morgan Fingerprinting. We use a 2048-bit long vector representation to introduce sparsity as well as accounting for the fact that a lot of the molecules in our index are pretty large with more than 200 atoms and 600 bonds. The query fingerprint vector is batch compared to the fingerprints in the index and a Tanimoto similarity score is obtained for each of the candidates.

Comparison

As shown in Figure 4.1, Tanimoto and graph similarity each have their own subtleties when searching for similar molecules. As Tanimoto similarity is computed through Morgan fingerprinting over the entire graph, it reflects the entire molecule's neighborhood through traversing the molecule in the generated sparse vector. Conversely, for sub-graph matching, any common substructure, no matter how small will result in a positive match. Each have their own strengths and we empirically show which one to use for which mode of search in Section 4.5.

4.3 Multi-Modal Search

Several methods were tried to fuse results between text and SMILES search. Some methods tried were simple conjunctive and disjunctive fusion of hits from BM25 and RDKit based searches as well as Reciprocal Rank Fusion [21]. We found that simple fusion methods did not work as intended because of the challenges of existing SMILES search models discussed above. Sub-structure search involves graph matching (i.e. boolean queries without match scores) which makes it difficult to informatively rank hits from this type of search. Tanimoto similarity based search does produce a bounded score between 0 and 1. However in the chemical domain, molecular structural accuracy is of paramount concern and Tanimoto similarity often does not have a direct positive correlation with information needs.

For instance, as shown in Figure 4.2, the enantiomers (containing different chiral bonds) of Thalidomide have a high Tanimoto similarity even though the fingerprinting method accounts for stereochemistry. Thus, a chemist searching for similar molecules to R-Thalidomide should not ideally be getting S-Thalidomide as a relevant hit.

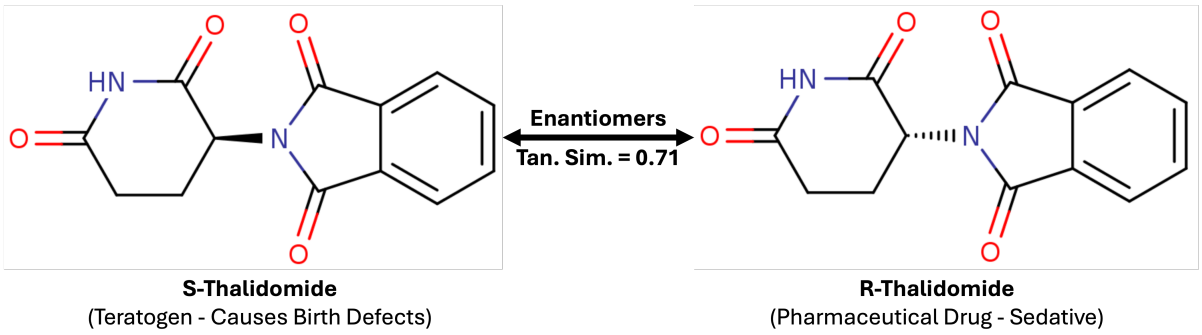

Figure 4.2: Shows how Tanimoto Similarity is high between the enantiomers of Thalidomide where the chemical properties of the R and S variant are vastly different, and despite the Morgan fingerprinting method accounting for stereochemistry. Similarly, both structures have almost all common sub-structures so sub-structure matching would also match. This shows the need to combine text and SMILES queries together which can produce more relevant candidates on searching for one or the other type.

Keeping this in mind, a reasonable first method developed was to re-rank passage hits from text search by passage hits from SMILES search and sorting the overlap in passages to produce the final ranking of passages. Thus in the above example, if a chemist included the text keyword "R-catalyst" along with the SMILES string for R-Thalidomide, our retrieval model would not likely rank any molecules similar to S-Thalidomide highly.

Thereby, given $R_{Txt} = \{(p_1, s_1), (p_2, s_2), \ldots, (p_n, s_n)\}$, the set of retrieved text passages with BM25-scores, and $R_{SMI} = \{(p_{s1}, s_{s1}), (p_{s2}, s_{s2}), \ldots, (p_{sn}, s_{sn})\}$ the set of passages matching a SMILES query through sub-structure search, the re-ranked candidate list $R^{'}$ is given by equation 4.1 as:

$$R^{'} = Sort(B \cup T) \tag{4.1}$$

$$B = \{\ (p_i, s_i) \mid (p_i, s_i) \in R_{Text} \wedge (p_i, s_j) \in R_{SMI}\ \} \tag{4.2}$$

$$T = R_{Txt} \setminus B \tag{4.3}$$

$Sort()$ sorts a set of text passage/score pairs of the form (p_i, s_i) by decreasing score. B refers to the set of passages that appear in both R_{Txt} and R_{SMI}. T refers to the set of passages in R_{Txt} without a SMILES sub-query match in R_{SMI}. This method is similar to the system used in [91] and ensures that the text candidates are considered as generally more informative along with minor improvements resulting from the SMILES search (compared to Text and math formula based

queries in [91]).

We acknowledge that there are limitations with computing a strict overlap between hits from the two modes of search. The first limitation is lack of chemically intuitive way to search the index where it is either through term matching (BM25) or graph/vector similarity (RDKit) that are independent of each other. The second limitation is a lack of joint context between the two modes of search, where one search does not inform the other mode or learn from it. We address these limitations in our follow up work in Chapter 6.

4.4 Query Generation and Relevance Assessments

We manually curate a set 142 PDF patents for creating the test collection and generating search queries. These patents have been carefully curated to only select patents that contain molecules related to one out of the 14 selected genes. These were compiled by a group of industrial chemists based on the topical relevance to drug discovery work. Table 4.1 shows a detailed distribution of the genes and the number of documents in each. In addition, it also shows the number of passages that have been collected along with the chemical entities detected that were successfully converted to SMILES. It was not guaranteed that every passage in the collection would be linked to a molecule diagram. For instance, out of the 2306 passages that were extracted for DGAT2, 116 of them had any links to diagrams. Some passages were also found have more than one link to molecules drawn in the PDF as shown in the rightmost column of Table 4.1, therefore were considered to be the more informative subset to draw queries from. There were 72 topics in total, 6 from each gene (12 genes in total) and each gene was assigned 2 queries of each modality — text, SMILES and multi-modal. Thus, the total test collection for the 12 genes comprised of 72 queries with 24 queries per type of modality.

Multi-Modal Query Generation Based on Linked Reaction-Only Passages

Queries were generated from the subset of paragraphs that had more than one link to diagrams but were evaluated using the entire test collection. The underlying principle towards generating the queries was that the majority of the molecules described in these patents could be considered structurally similar but can be chemically very different in terms of exhibited properties like binding affinities, synthesis process, etc. Some reagents and substrates might be so common that they are used in the synthesis of multiple compounds. This is because patents from different pharmaceutical

Table 4.1: Patent PDF Test Collection Metrics. Topics for the collection are designed using the passages and diagrams extracted from these 131 PDFs. There are 6 topics for each gene; 2 for each mode of search — text, SMILES and Multi-Modal. Thereby, the 72 topics created represent a diverse set of information needs categorized by gene type.

Gene	Num. PDFs	Total Pages	Passages with SMILES	Total SMILES in Passages	Passages with Single Linked Diagram	Passages with Multiple Linked Diagrams
CD73	14	3682	8584	14777	929	963
DGAT2	12	1279	2306	5223	74	42
DKGa	1	465	242	373	19	73
GLP-1	9	2997	2926	4862	283	346
Helios	2	599	1089	2115	65	46
IRAK4	35	6685	11059	20740	1799	1146
KHK	3	350	175	247	23	0
KRAS	1	769	2133	3892	37	56
Mcl1	1	311	1531	2741	25	0
PARP7	2	697	1908	3274	194	241
PRMT5	1	632	1509	2082	181	116
TEAD	50	13835	23717	49288	3187	4360
Total	**131**	**32301**	**57179**	**109614**	**6816**	**7389**

companies that target the same gene develop molecules that usually have common sub-structures but different overall structure. This is because these molecules usually contain a certain atom group which is the main part of the molecule that binds to a gene protein to inhibit its activity. In a retrieval context, this also makes the problem harder because a lot of drawn molecules and their synthesis pathways described in text will share a lot of similarities between them, structurally or chemically, and in terms of reaction specifics like temperature, yield, etc. Thus, the motivation was creating a test collection where searching for a specific chemical information out of deceptively similar ones can be quick and effective using flexible queries in more than one modality.

Table 4.2 shows the queries generated for each of the 12 genes in the test collection. The queries have been specifically designed to simulate information need for chemists in drug discovery. The majority of text queries are asking about either a reaction given a reactant or a synthesis pathway to obtaining a specific product. This aligns well with the information needs in drug discovery where chemists tend to look for different synthesis pathways of specific functional groups. Furthermore, all the SMILES queries refer to specific sub-structures of potential interest and not full molecules. This is motivated similarly by often ambiguous searches by chemists to find synthesis pathways of interest for a sub-structure that is known to bind to certain proteins in genes and inhibit their activity.

4.4.1 Relevance Assessment

The queries were created keeping in mind that human search queries can be imprecise and each molecule that was being referred to using the query would be found only once in the entire collection thereby giving us a measure of exact high relevance. Our process of assigning relevance score to the document set for each query was completely automated and based on the gene-based chemical significance described above.

In other words, for a particular query, that referred to a specific passage in a specific PDF belonging to a specific gene, any candidate that was from the same PDF was considered to be of medium relevance. The justification for this is that if the user is at least pointed to the correct gene and the correct PDF of the gene, they are still close to their exact information need. Similarly, any candidate that pointed to a passage in a different PDF but of the same gene was considered to be of low relevance. Our levels of relevance are formally defined below:

- Same Gene, Same PDF, Same Passage – **High Rel. (3)**

Table 4.2: Text and SMILES queries generated for each of the 12 genes in the collection – 72 topics

Gene	**Text Query**	**SMILES Sub-Structure Query**
KHK	difluoromethyl pyrimidine obtained with 400MHz NMR	[H]C(F)F
	Preparation of cyano methylazetidin acetic acid from triflouromethyl pyridine carbonitrile	C1CN=CN1
DKGa	flouroquinazoline along with flouroaniline	C1CC=NCN1
	quinazolin benzoxazepine synthesized from a form of cyclopropane	FC1=CC=CC=C1F
PARP7	used bromo-flouro instead of 2H-isoquinoline	FC1=CC=CC=C1Br
	synthesis of flourooxetan pyrimidine	ClC1=NC=CC=N1
DGAT2	preparation of pyrrolidin-methanone	N1C=NC2=C1N=CC=C2
	preparation of pyrimidine carboxylic acid	C1CCNCC1
Helios	preparation of intermediate piperidine dione	O=C1NCC2=CC=CC=C12
	preparation of tert-butyl cyclohexyl carbamate	NC1CCCCC1
IRAK4	Synthesis of pyrimidine carboxamide	CC1(C)CNCCO1
	preparation of picolinamide from triflouromethyl pyridin and amine	N1C=NC2=C1N=CC=C2
KRAS	intermediate compound as flourodihydro cyclopropane pyrrolizin	C1CC2CCCN2C1
	2-(7,8-difluoro-3-(methoxymethoxy)naphthalen-1-yl) 4,4,5,5-tetramethyl-1,3,2-dioxaborolane	FC1=CC=CC=C1F
CD73	production of imidazo pyridazine after triturating with water	C1=CN2N=CC=CC2=N1
	Starting with a solution of isobutylimidazo pyridazin amine	CC(C)(C)OC1=NC(OC(C)(C)C)=NC=C1
Mcl1	synthesis using dichloro pyridazine	N1C=C2C=NC=NC2=N1
	mixture of pyrazole carboxylate and palladium dichloride in a substrate of ethyl acetate	CN1C=CC=C1
PRMT5	picolinaldehyde	CC(F)(F)F
	preparation of N-acetyl-amino pyrazolo quinoline carbohydrazide	CN1N=CC2=C1C1=C(C=CC=C1)N=C2
TEAD	dichloromethane and pyridin acrylamide	FC1=CC=CC=C1
	azetidin methanol hydrochloride used as a catalyst to produce pyrazolo pyridin methyl acrylamide	C1CNCCN1
GLP-1	cyclopropyl pyrazol oxy flourobenzonitrile prepared with piperidin carboxylate	C1CCNCC1
	Intermediate methylcyclopropyl amino benzoate for substituting methanamine	NC1=C(N)C=CC=C1

- Same Gene, Same PDF, Diff. Passage – **Medium Rel. (2)**
- Same Gene, Diff. PDF, Diff. Passage – **Low Rel. (1)**
- Diff. Gene, Diff. PDF, Diff. Passage – **No Rel. (0)**

The major limitation of the test collection is the lack of expert evaluation before or after generation of the topics. While we believe in the efficacy of the relevance levels and the assigning of relevance based on gene, we acknowledge that the queries themselves may not represent or closely relate to actual information needs of drug discovery but rather follow the overarching pattern of searching for sub-structures without being specific. In Chapter 5, Section 5.2.4, we propose a method by which all query hits will be manually evaluated by a domain expert.

4.5 Preliminary Results

Our experimental methodology was guided by the intuition that for CIR where a lot of closely related information can be present, combining search results or using one result to guide the other result can be helpful. The experiments show that the best search method can be dependent on how precise the user can be with their queries and whether they can provide more than one modality of information. For instance, searching for a full compound name along with it's reaction conditions can lead to better results than a vaguely formulated query with only a few matching sub-groups of a chemical name. Furthermore, combining a text query along with a full molecule structure will lead to more relevant results than a smaller sub-structure SMILES query. However, this might seem counter-intuitive with the motivation of enabling chemists to search with small sub-structures of interest. We dive deeper into modality specific search next. We start with a brief report on the performance of our molecule diagram parser MolScribeV2 [33] and then present the preliminary sparse search results.

Diagram Parsing Results

For evaluating our diagram parser MolScribeV2, a custom annotated set of 1832 molecules was created from relevant patent documents with expert annotated molecule SMILES. Our training dataset consisted of 5M million unique molecule SMILES collected from PubChem [60], Zinc250k [2], ChemBL [94], MOSES [113]. These molecules were converted to synthetic images at training

Table 4.3: Comparison of MolScribe and MolScribeV2 on a set of 1832 molecules. Numbers in parenthesis indicate the number of molecules which could be successfully computed for the metric.

System	Accuracy(%) SMILES Match ↑	Leven-shtein ↓	Tanimoto Similarity ↑	Graph Edit Distance ↓
MolScribe (Baseline)	88.02	34.89	0.976 (1761)	0.039 (1666)
MolScribeV2 (5M Train)	90.31	3.27	0.980 (1790)	0.032 (1715)
+ Enh. Pos. Emb.	90.09	4.30	0.982 (1790)	0.033 (1717)
+ Crop/Add Margins	**94.61**	**0.83**	**0.987 (1834)**	0.035 **(1783)**
+ Doc. Degradation	94.17	1.52	0.983 (1813)	**0.012** (1756)

time using the Indigo toolkit[1]. For more details on the training method, please refer to Section 3.2. We specifically do not report on performance on benchmark datasets for diagram parsing because our custom dataset consisted of harder molecules. Furthermore, the public benchmarks only report exact match accuracy whereas we provide more detailed and informative metrics like Levenshtein, Tanimoto Similarity and Graph Edit Distance.

Table 4.3 shows the improvements of MolScribeV2 compared to the original MolScribe. Accuracy represents the exact SMILES match between the ground-truth and the predicted. Levenshtein is a string based distance metric which computed the number of additions, substitutions and deletions to make the predicted SMILES match with the ground-truth SMILES. Tanimoto Similarity is computed using the vector based Intersection-Over-Union (IoU) match between the Morgan Fingerprints of the predicted molecule SMILES and the ground-truth SMILES. Graph edit distance is the number of additions, deletions and substitutions required in the predicted molecule graph with atoms as the nodes and bonds as the edges to match the ground-truth graph. The performance is seen to gradually improve with each additional improvement on top the baseline MolScribe. We see the greatest improvement in scores by the addition of cropping and adding margins to the training data through an exact match accuracy of 94.61%. This is substantiated by the fact that molecule detection models that predict the spatial location of molecule diagrams in a PDF image are not perfect with respect to their position and size and often add additional graphics such as margin borders or snippets of a different molecule diagram or unrelated text. It is also found that adding additional noise such as salt and pepper noise and gaussian noise did not help improve the overall performance and reduced slightly to 94.17%. We attribute this behavior to the nature of chemical diagrams where a lot of implicit carbon atoms are placed at the junction of two lines and

[1] `https://lifescience.opensource.epam.com/indigo/`

Table 4.4: Baseline Results: Hybrid Sub-Structure Multi-Modal Search. Relevance Level 1 includes hits belonging to PDFs addressing the target gene, Level 2 includes passages from the same document as the target passage, while Level 3 includes only the target passage.

Metric	**rel≥1**	**rel≥2**	**rel=3**
Precision			
P@1	**0.708**	0.583	0.167
P@5	**0.600**	0.525	0.049
P@10	**0.510**	0.446	0.029
nDCG			
nDCG@1	**0.487**	0.444	0.167
nDCG@5	**0.528**	0.492	0.214
nDCG@10	**0.488**	0.453	0.226
MRR	**0.763**	0.644	0.218

if that junction becomes too noisy, there might be additional carbon atoms predicted in the output which reduces overall performance. Overall, the greatest improvement came from the addition of positional embedding at the input as well as the augmentation of addition/deletion of edges of synthetically generated training images.

Retrieval Results

We demonstrate our BM25 and RDKit based search performance using the test collection of 72 topics created as described in Section 4.4. Our experimental methodology was guided by the intuition that for CIR where a lot of closely related information can be present, combining search results or using one result to guide the other result can be helpful. The experiments show that the best search method can be dependent on how precise the user can be with their queries and whether they can provide more than one modality of information. For instance, searching for a full compound name along with it's reaction conditions can lead to better results than a vaguely formulated query with only a few matching sub-groups of a chemical name. Furthermore, combining a text query along with a full molecule structure will lead to more relevant results than a smaller sub-structure SMILES query. However, this might seem counter-intuitive with the motivation of enabling chemists to search with small sub-structures of interest. We dive deeper into modality specific search next.

Table 4.5: Detailed Performance Results for search modes at relevance levels 1 and 3. SMILES-only search has a larger degradation in performance due to common sub-structures in the collection, and the SMILES queries defined as sub-groups from molecules of interest. (***Baseline model**)

	P@1		**P@5**		**nDCG@1**		**nDCG@5**		**MRR**	
Query Type	**rel≥1**	**rel≥3**	**rel≥1**	**rel≥3**	**rel≥1**	**rel≥3**	**rel≥1**	**rel≥3**	**rel≥1**	**rel≥3**
Text	0.583	0.125	0.583	**0.058**	0.403	0.125	0.497	**0.216**	0.678	0.204
SMILES (Tanimoto)	0.292	0.042	0.208	0.008	0.194	0.042	0.176	0.042	0.382	0.042
SMILES (Sub-Structure)	0.208	0.0	0.175	0.008	0.139	0	0.154	0.026	0.23	0.023
Text + SMILES (Tanimoto)	0.542	0.083	0.575	**0.058**	0.361	0.083	0.482	0.20	0.656	0.183
Text + SMILES (Sub-Structure)*	**0.708**	**0.167**	**0.60**	0.049	**0.487**	**0.167**	**0.528**	0.214	**0.763**	**0.218**

(1) Multi-Modal Search: Table 4.4 shows the baseline scores at all relevance levels 1, 2 and 3. The multi-modal search was conducted through two different ways of SMILES search as discussed earlier and was used to re-rank the candidates of the text search. In our experiments we see a model increase in overall performance when we used sub-structure based multi-modal search compared to the text only search. However, conversely we see a drop in performance when Tanimoto Search was used and degraded the existing text results. We achieved the highest MRR at 0.763 in this mode. This behavior was in contrast to when SMILES search was used in a standalone fashion and Tanimoto performed better than sub-structure. We think this is because even though sub-structures had fewer relevant hits higher up, the hits when relevant were always of a very high relevance (relevance score of 2 or 3).

Figure 4.3 shows an example of how through combining sub-structure based results with text for a particular query, the system could correct an irrelevant text based top hit to a relevant one. This figure also further demonstrates the challenges with using sub-structures SMILES as queries. Even though the ground-truth text match for that query is `2,4-dichloro-5-(3-fluorooxetan-3-yl)pyrimidine`, the SMILES search matched a related structure in the same page with the name `2-chloro-5-(3-fluorooxetan-3-yl)pyrimidin-4-amine`. Both of them are technically valid hits as both contain the query substructure in them.

In Table 4.6, the multi-modal search clearly shows that there are no significant differences between re-ranking and keeping the text results. We attribute this behavior to a combination of the query sample size being small and the lack of a better way to fuse results. We also think that the chosen sub-structures in and of themselves struggle to provide relevant signals due to the huge noise accumulated from many different compounds extracted from different genes in the collection.

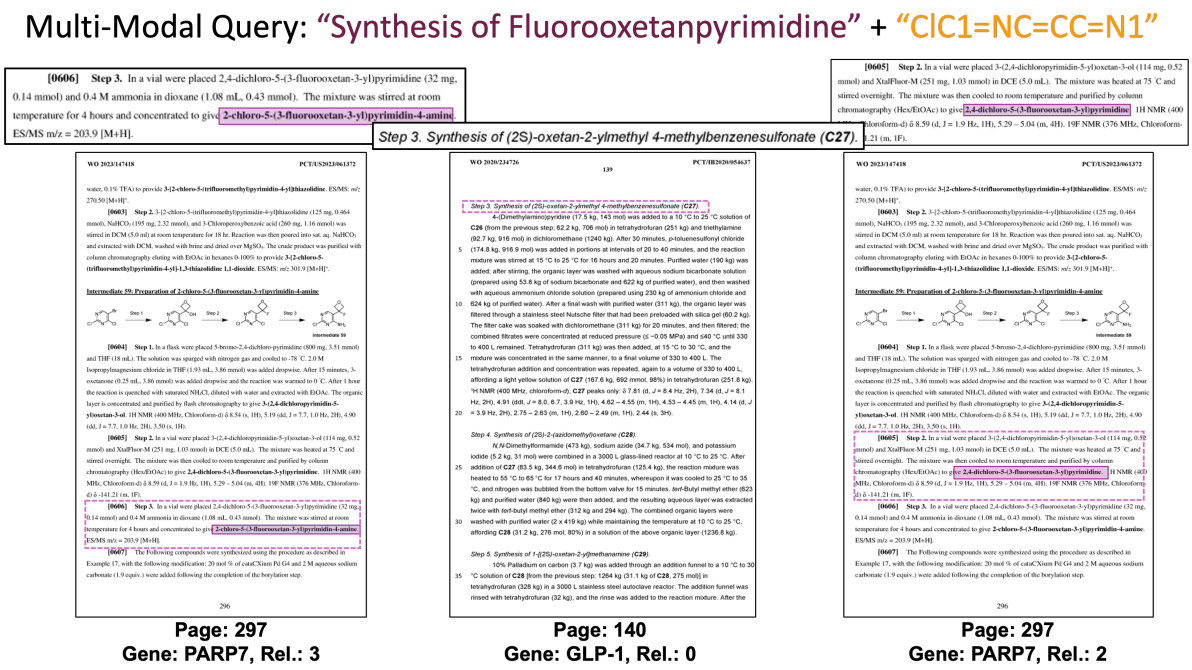

Figure 4.3: Example Results for Query Using Multi-Modal Search (Text + SMILES). In this example re-ranking text searches by sub-structure SMILES matches improves results. **Left:** The ground-truth of the given query. **Middle:** The top hit from Text Only Search. **Right:** Top hit after re-ranking using the SMILES search results. The top hit from text search led matched with a different gene PDF due containing similar word tokens — "synthesis", "of" and "oxetan". This was rectified by reranking using the SMILES search results where the users were led to a passage at the same page as in the ground-truth.

On additional hypothesis testing, we found that overall, out of the 24 unique topics, Multi-Modal search improves P@1 and nDCG@1 on average across the different relevance thresholds for 3 queries and degrades the performance of one query.

Next, we aim to compare the baseline scores with other modes of search by selecting only text or SMILES query standalone and comparing with the baseline scores. Our queries have been designed to be reflective of the complex need of chemists in drug discovery, the SMILES queries have been intentionally made of sub-structures which are commonly found across the entire collection. The compound names mentioned as text are also incomplete chemical name mentions. The objective for this section to see how much the performance degrades by removing any one of the modes.

Table 4.6: Bonferroni Corrected p-values vs. the text-only BM25 baseline. Statistically significant differences have been marked with a dagger ($p < 0.05$). Items in parentheses indicate retrieval model used for SMILES queries by query type.

Metric	Query Type		
	SMILES (Sub-Struct.)	**Multimodal (Sub-Struct.)**	**Multimodal (Tanimoto)**
rel ≥ 1			
P@5	†0.003	1.0	1.0
nDCG@5	†0.01	0.92	0.71
rel ≥ 2			
P@5	†0.021	1.0	1.0
nDCG@5	†0.029	0.88	0.66
rel $= 3$			
P@5	0.122	1.0	1.0
nDCG@5	0.123	1.0	1.0

(2) Text Only Search: The text queries in our collection have tried to emulate the type of queries chemists would like to use. It can be challenging for a user to know the exact compound they might be looking for or the exact reaction condition that might go with it. Therefore our queries have been intentionally incomplete where only certain sub-group names have been used instead of the entire chemical formula. Table 4.5 shows the result of text search at relevance cutoff thresholds of 1 and 3. It shows an expected pattern of lower metrics of both P@X and nDCG@X when the relevance threshold becomes more strict. The interesting observation was that at relevance cutoff of 3, it still gave highest proportion of relevant results among the top 5 ranks. We attribute this behavior to our chemically relevant tokenization scheme where the possible of word matches in the BM25 index is increased and has a certain level of tolerance if someone accidentally misspells a chemical name. We found this mode of search to be stable enough to use it as our baseline for statistical testing for the other modalities of search.

(3) SMILES Only Search: As shown in Table 4.5, both methods of searching through SMILES strings (Tanimoto Similarity and Sub-Structure) perform the worst out of three types of search. This was an expected result as our SMILES index has about 13791 unique SMILES and the queries are composed of only sub-structures. This means, even though the returned results might be

"chemically relevant" they do not meet the information need given the query.

Table 4.6 shows the Bonferroni corrected p-values taking the text search results as a baseline. The SMILES search results indicate that the distribution is highly likely to be much different than the results obtained from text mode. This is again expected as there might be same sub-structures of interest across multiple genes and many passages not related to the ground-truth may have them. Furthermore, in our experiments with SMILES search, we limit the substructure search to Top-1000 results, and Tanimoto search to Top-20. This is standard practice for evaluating collections where only a subset of candidates are expert annotated. We will explore the impact of this thresholding in future work.

One interesting observation from Table 4.6 is that the p-value increased with stricter relevance constraints going from statistically significant to insignificant. We attribute this behavior to our query set being relatively small (24 different topics per modality).

We believe that this initial sparse keyword and graph matching based search is a good first step to create a test collection for the retrieval community. It comprises of a simple system to generate a baseline and standard ways to fuse results from two different ways of search. More work will be needed in future to come up with better ways of fusing the two modalities of search and increasing the set of queries for better statistical testing. Furthermore, the current system lacks the ability to infer joint context from the text and SMILES modality. We provide this baseline as a starting point to the community to develop tools for multi-modal searching of chemical documents. These limitations are addressed in the proposed work section (Chapter 6). Furthermore, this modeling approach will serve as an important benchmark for comparing the new proposed dense model, providing important performance metrics that distinguish sparse from dense models for a chemical information retrieval task.

4.6 Summary

In this Chapter, we first defined the created index pattern that allows scalable and fast search and lookup of passages, diagrams and their associated metadata such as the SMILES strings, IUPAC names, PDF pages and spatial locations as bounding boxes within those pages. This was followed by a detailed discussion of the currently developed modes of sparse searches where BM25 was used for textual queries, RDKit was used for SMILES queries and a combination of the two was used for multi-modal text and SMILES queries. The limitation of these modes of search was briefly

discussed that showed how a lack of contextual representations limit the perceived performance of the designed baseline methods. An expanded discussion of the developed search interface followed which talked about the motivation and utility of in-context PDF-based chemical search. The test collection to evaluate the sparse modeling approach was then introduced that demonstrated the strengths and weaknesses of the proposed modeling as well as the topic generation methodology. Finally, a discussion followed on the analysis of the preliminary baseline experimental results on the generated test collection.

The next Chapter will introduce the improvements to the limitations of modeling approach as well as methods associated with extracting the indexing data and generating the test collection.

Chapter 5

Test Collection and Expert Assessment

This chapter introduces and describes the methods and processes surrounding the creation of the chemical expert curated data collection and improvements in the extraction and indexing of raw chemical text passages and its corresponding drawn molecule diagrams and reactions. The extraction work improves the related technical improvement goals T1, T2, T3 described in Section 1.2.2. The resulting relevance graded test collection comprising of 35 expert curated and annotated diverse queries from the extracted data answers the first research question RQ1 on the need of involving chemists in the query creation and assessment process as outlined in the following sections.

To ensure better coverage of context in the patent PDFs, we changed the text extraction process by replacing Tesseract with the neural-based Surya OCR. This increased the number of extracted passages. We also describe how adding an additional chemical named entity dictionary allowed us to identify and convert more text-based chemical entities within these passages into their corresponding SMILES representations. One of the important features to improve search quality and better incorporate chemists' information needs was searching for passages and drawn molecules and reaction pathways together. For this, we developed a pooling method that combined results from linked passages and diagrams using bucketing and re-ranking as described in Section 5.2.2.

The pooling process for generating retrieval candidates for expert evaluation stage went through a series of iterations. This was to make the evaluation process simple and convenient for chemical experts. This involved careful and honest conversations, in a collaborative process that recognized

the importance of understanding cross-domain gaps in expertize, limitations, strengths and common pitfalls. We describe the efforts in Section 5.2.1 and how a single query annotation time was reduced from over 3 hours for 10 candidates to under 30 minutes for 20 candidates.

5.1 Improving Chemical Named Entity Extraction and Conversion

The extraction process of text passages and associated chemical entities and creating the associated SMILES was improved. This allowed for more coverage of relevant information with documents that were earlier missed due to recall arising from leakage during the passage segmentation, OCR and OPSIN conversion of entities to SMILES. We show in this section how changing the extraction and conversion process led to a significant increase not only total passages but also quality of data. This extraction change also led the groundwork to include additional chemical process based passages that were either too sparse or noisy to be sampled before.

Unifying Passage Segmentation and OCR into a single end-to-end pipeline Instead of using LayoutParser [130] as the passage segmentation model and Tesseract [56] as the OCR on the segmented passages we transition to use a unified segmentation and text OCR pipeline called SuryaOCR [109]. This CNN-based neural model extracts passage regions and then runs OCR on the passages in a unified pipeline. As shown in Figure 5.1, segmentation quality is is significantly improved by avoiding common previous issues like under-segmentation or over-segmentation that led to either lines of a paragraph being missed or the same passage split into many segments. Furthermore, the OCR quality was also significantly improved. The downstream chemical named entity system, ChemDataExtractor [92], was extremely sensitive to errors in the OCR stage. This meant simple and common Tesseract errors that are commonly found in IUPAC names like “1” or “i” were less likely to be confused and led to more CNER coverage. This eventually increased the SMILES coverage from passages as well.

Increasing SMILES conversion and linking through external knowledge In addition to changing the initial passage extraction and OCR stage, we realized early on that a lot of molecules were being missed by OPSIN [82] when presented with the detected CNERs from ChemDataExtractor [92] stage. It was discovered that even though ChemDataExtractor was good at identifying all different types of chemical entites with high recall – IUPAC names, com-

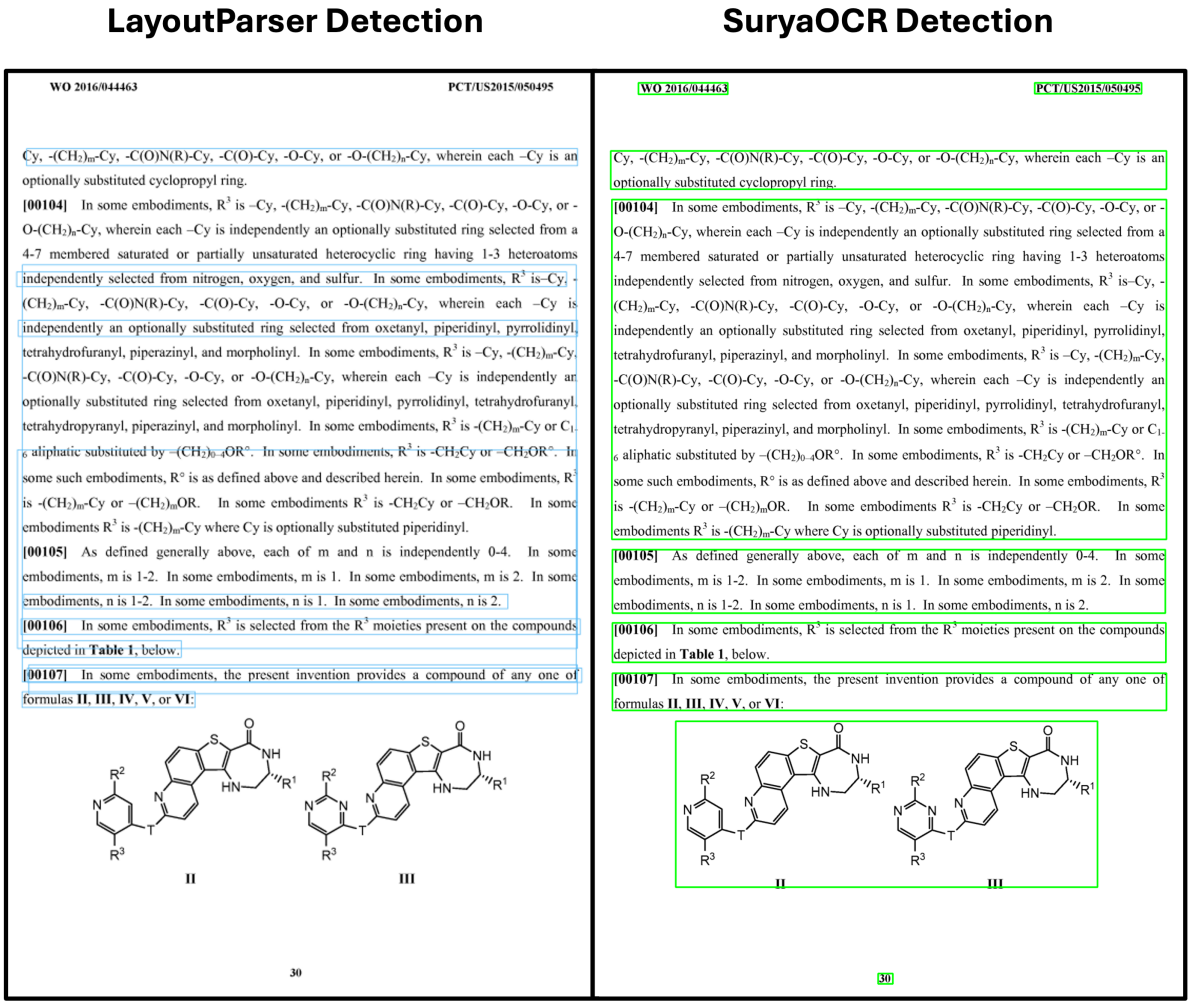

Figure 5.1: Comparison of passage detection performance between the currently used LayoutParser and the proposed change to SuryaOCR detection. Compared to LayoutParser, SuryaOCR has much better recall for paragraphs, does not have overlapping detections, does not miss parts of passages and fits tightly around the passage margins.

mon names and industrial names, OPSIN was only designed to convert perfect IUPAC structured names into their corresponding SMILES. This meant that if a chemical entity with the IUPAC: `N-(4-hydroxyphenyl)acetamide` was being referred with its corresponding common name such as acetaminophen or the industrial name (Tylenol), it was being missed by OPSIN. This posed a challenge as pharmaceutical patents (and chemistry research articles) commonly abbreviate their molecules or use synonyms that are common to identify by trained chemists manually reading them.

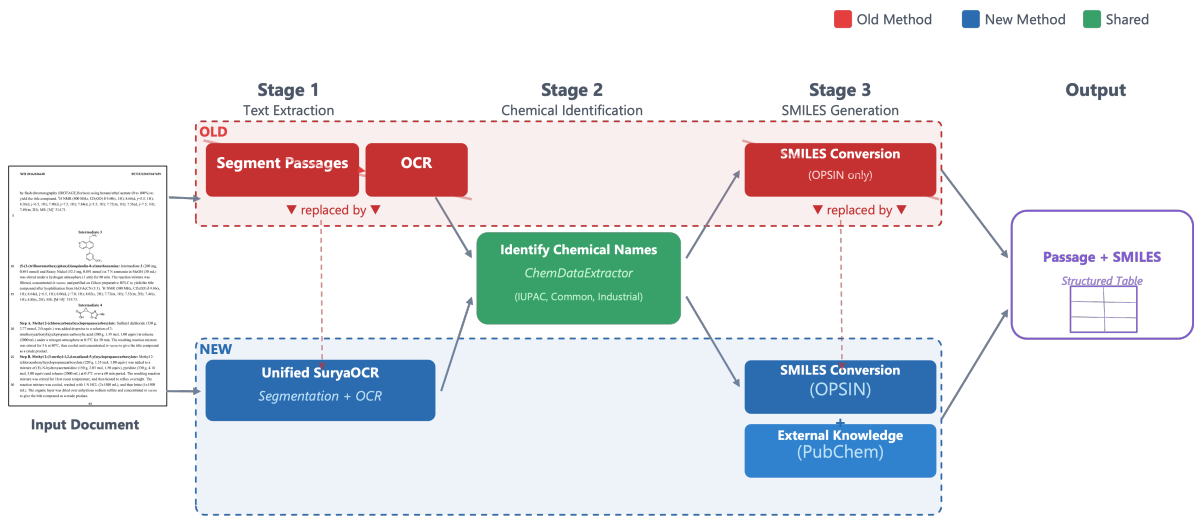

Figure 5.2: Full Data Extraction Pipeline Overview: Unified Segmentation and OCR pipeline was used to extract passages and the CNERs detected by ChemDataExtractor went through a tiered conversion attempt process through OPSIN and then PubChem lookup databases. The final extractions of passages and any linked SMILES were indexed into the raw index with its metadata. Refer to Appendix B describes the indexes.

To mitigate this and increase the efficacy of data extraction and generation process, we incorporate an external open knowledge database from PubChem [60]. PubChem contains a collection of 121M molecules along with all their metadata such as chemical and physical properties, reactivity, melting points etc. More importantly, each molecule also contains their corresponding IUPAC names, SMILES representation as well as any synonyms comprising of IUPAC variations, industrial names and common names. Using the PubChemAPI, we curate collections of fast lookup molecule databases. The descriptions of these databases are mentioned below.

1. **IUPAC Variations → Official IUPAC (170M):** To capture any variations in IUPAC Names and map it to the official IUPAC Name
2. **Common/Industrial Names → Official IUPAC (171M):** To capture common and industrial names and map it to the official IUPAC Name
3. **Official IUPAC Name → Canonical SMILES (77M):** To map any official IUPAC Name to is corresponding SMILES

Using this 3-tiered key value lookup databases, external knowledge was meaningfully used to in-

Table 5.1: Total Passages and Linked SMILES extracted from the 141 chemical patent PDFs containing 32,301 pages. Increase in total passages is the impact of the segmentor and OCR unification while the increase in the number of valid SMILES linked is the impact of the external knowledge base.

Metric	**Before**	**After**	**Increase**
Total passages extracted	8,731	377,904	42x
Total passages with linked SMILES	2,779	373,107	134x
Total unique SMILES	585	32,266	54x

crease conversion rate of CNERs detected from the ChemDataExtractor stage. All CNERs were first passed through OPSIN. This step is the same as before. Any SMILES that failed conversion were then passed through the lookup dictionaries. These key-value databases provided fast lookup times given a key without loading all the values into memory. RocksDB [35] was used for the orchestration and memory management of the databases. Any IUPAC variations if detected were converted to its full official IUPAC name. Same for the common/industrial names. Any and all IUPAC names were now used as the keys to lookup the corresponding SMILES. Figure 5.2 shows the overall extraction architecture.

This was an important contribution towards our final research goal of not only the creation of a novel chemical expert graded relevance test collection but also to harvest enough training data for our neural training model. Table 5.1 shows the impact of our changes on the generated collection compared to the preliminary work. The change in the extraction stage into a unified segmentation and OCR framework through SuryaOCR led to a 42x increase in the detected passages while incorporating the external knowledge base from PubChem led to a 54x increase in the total number of unique SMILES that were linked to passages.

The next step involved utilizing this collection to generate the initial pooled candidates for expert assessment. The process of expert assessment involved developing a pooling algorithm to generate the initial candidate sets on the technical side. Majority of the work involved collaboration with chemical experts to define a graded relevance criteria for the assessment process and developing a convenient visual augmentation for experts to score raw candidates by avoiding repetitive work and fatigue.

5.2 Creating the Graded Relevance Test Collection

The pooling process was one of the most complex processes within this work. This involved gathering the best experts for the sub-domain of organic and synthetic chemistry, defining a novel chemistry inspired relevance assessment criteria, aligning with chemists on the major objectives and computing processes that were needed to accomplish the task and creation of the annotation mechanism to enable experts to annotate queries with the least amount of wasted effort and time spent. The entire process involved a 5-month process of iterations adjusting the entire collection process from redefining the pooling process, models used in the pooling, modalities used as well as adjusting the relevance criteria. In this document, we only describe the final collection generation process for brevity and flow.

5.2.1 Alignment with Expert Chemists

The nature of our proposed test collection required a group of chemists who were specialized within a very specific domain of chemistry research – organic and synthetic chemistry with a focus on pharmaceutical small molecule chemistry. For this purpose, we reached out many research groups with closely related research areas and went ahead with two research groups – one from Rochester Institute of Technology (RIT) and the other University of Rochester (UR).

The expert group consisted of three university professors and three chemistry students. All of the students had a chemistry background and were guided by the respective professors. While one of them was an advanced undergraduate student, the other two were senior PhD chemistry students. The initial month was spent making the retrieval task clear and how their expertise was essential towards the creation of the collection. This included laying out an initial plan to create the topics for assessment, coming up with a draft relevance criteria for scoring candidates, and what modalities should be ideally scored. This followed months of training and aligning with the experts on what the annotation process would look like and whether the candidate evaluation process makes sense to both parties.

5.2.2 Pooling Process

Refining the pooling process required a series of conversations with the chemists about what kind of information needs are relevant when searching for information within pharmaceutical patents

and how the candidate generation process could meaningfully be fine-tuned to fit those criteria. In a retrieval sense, it was figured out that information needs can sometimes not closely match the query terms or could be unlike a traditional keyword-based search. The major point of realization was to understand that:

1. Simple keyword-based searches within chemical literature already existed with Reaxys and SciFinder and just page-level results of candidates as an additional feature was not enough to convince chemists of the usefulness of the work. This meant a more semantic level search with a retrieval model understanding the implied chemical constraints was required.

2. Chemists wanted more targeted search across the archive of documents. A lot of the times, chemists wanted to be directly pointed towards a reaction schema as a whole rather than a single reaction or a single intermediate molecule synthesis process. This meant just treating the text passages as valid candidate hits was not enough.

3. Frequently, for certain information needs, the full reaction schema definition or synthesis process was not available within a single passage, in fact, not even a single page. This meant considering only a single-page passage result as an independent candidate did not fully satisfy the information required.

4. Scoring with only candidates recorded in a spreadsheet tool along with an encoded SMILES string and text query was too prohibitive for a chemist given their time constraints. This meant the assessment process had to be simplified to make the evaluation process as convenient as possible.

To address these concerns, a new pooling process was developed. The pooling process joined both the text and diagram results as a single retrieval unit. To treat treat text and molecule diagrams as a single retrieval unit our pooling approach was a two pronged one. All diagrams were represented as a SMILES unit for retrieval, similar to the SMILES converted from IUPAC names within text passages. MolScribeV2 [32] as discussed in Section 3.2 was used to extract and index all diagrams in the documents. This created two unpaired indexes of text passages and molecule diagrams. Text passages and their corresponding CNERs converted to SMILES were then linked with diagram matches. However, to ensure both passages and diagrams could be used for efficient retrieval, there were two main factors to consider.

Many extracted passages did not have any linked diagrams. This was because either (1) the passage actually did not contain any CNERs that were explicitly drawn within the PDF or (2) the text

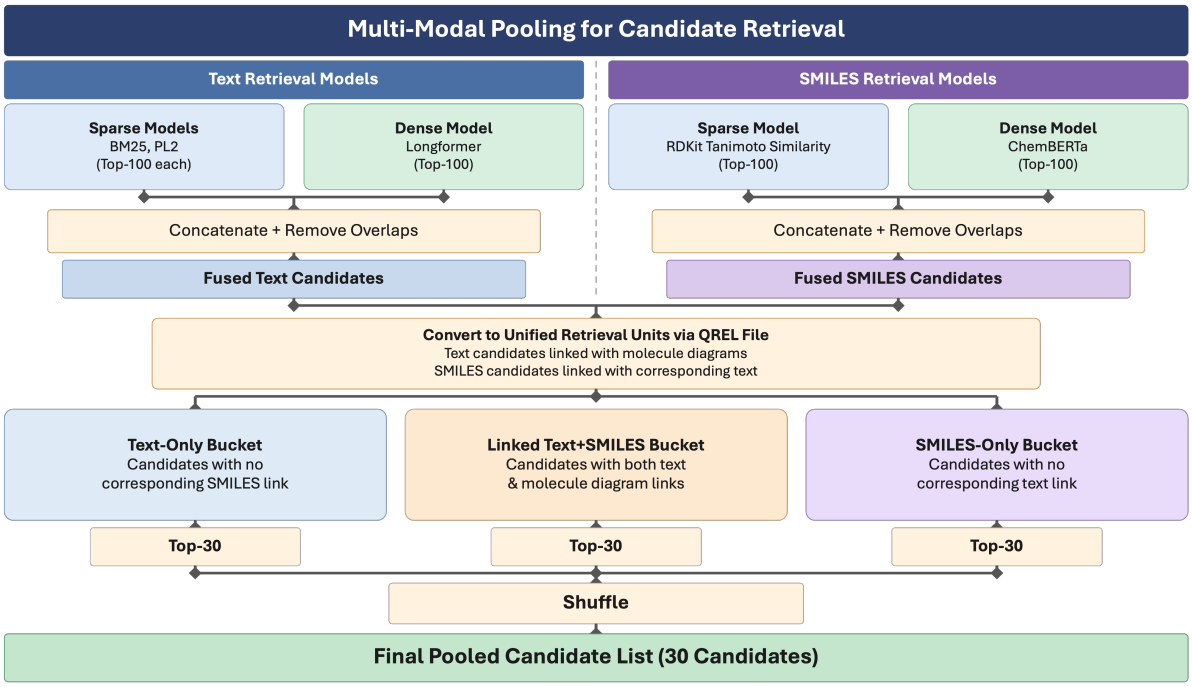

Figure 5.3: Candidate Generation Process for pooling Top-K candidates for human annotation towards creating the expert annotated candidate subset for each multi-modal query

CNERs failed to be converted to SMILES or (3) there were conversion errors from the molecule diagram parser. Taking all of this into account, the best path forward to get enough training data for the pooling models was chosen as using only the text CNER SMILES linked to the passages as the positive pairings. This presented us with 300K passage-SMILES pairs to be used as training data. The implicit assumption behind this choice was that as the diagrams were converted to the same downstream SMILES representation, any retrieval on the diagram index using SMILES would work equally as well as the CNER SMILES index.

The other factor was deciding the correct indexing strategy. Text and SMILES inherently belong to two separate modalities and their indexes had to be independent. This was a design choice to ensure existing off-the-shelf SMILES and text retrieval models work without a lot of implementation effort. Figure 5.3 shows the overall cadidate pooling process. Similar to other benchmark collection generation methods like the TREC competitions [24, 22, 25, 23], TREC-COVID dataset [123], BEIR [140], MIRACL [162] and Cai et al. [10], we use 3 independent models for chemical passage retrieval – BM25, PL2 and LongFormer [79]. For SMILES candidates retrieval, we use RDKit based Tanimoto Similarity search model and ChemBERTa [18]. The candidates from each of the models were fused together using QREL file that contained the list of all text and SMILES candidates

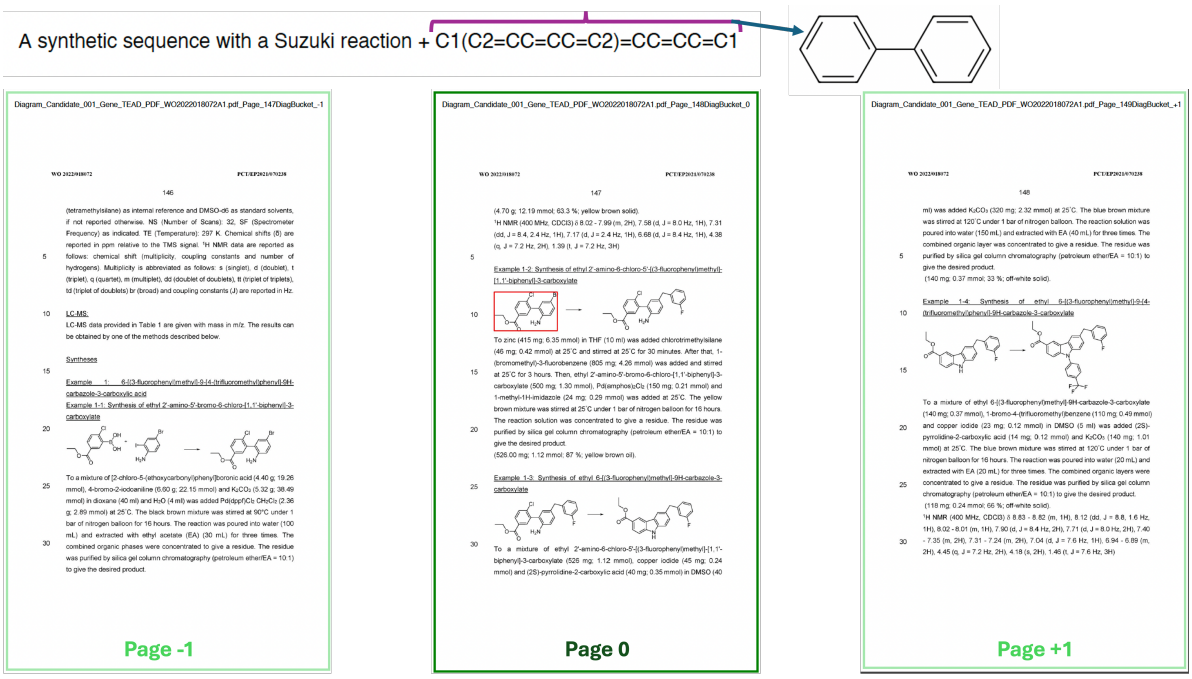

Figure 5.4: The title page and a diagram candidate pages excerpt from a pooling query. The title page has the actual text query marked as well as the SMILES query visualized which removed a lot of mental effort for chemists to convert an encoded SMILES into its 2D-structure. The 3 pages depict the actual candidate hit in Page 0 along with its full page context. The Pages -1 and +1 add additional neighborhood context for judging the candidate relevance. **Note** that each candidate triplet can be retrieved from any of the 141 PDFs and the 32,000+ pages within the dataset collection.

as well as if any text passages had links to any diagrams and vice-versa. This created 3 separate buckets to pool from. The final candidates after pooling were picked randomly from the three types of candidate buckets to ensure diversity and an even spread between different modalities of candidates. This pooling process addressed the chemists two main concerns – using an additional modality of querying except for generic text and having both passages and chemical diagrams as valid retrieval candidates. This pooling style also fused dense models along with sparse models to cater to different information needs of simple keyword matches (BM25, PL2) as well as matches that need more chemical knowledge (LongFormer, ChemBERTa).

The final requirement of the chemists was to create a less prohibitive structure and flow towards scoring candidates. A new candidate list PDF was created. This list contained all the candidate passages and molecules embedded within their original pages (in-context). Furthermore, it also

contained the context of the page before and after where the actual text or diagram was found. This created an extended context for assessors to judge the candidates. This became really useful as we created our relevance definition criteria in consultation with our expert chemists. The candidate PDF contained the final pooled list from the pooling stage where the exact passages and molecules were boxed in as well as their surrounding context to enable chemists to make a more informed decision about the relevance of a particular candidate. Figure 5.4 shows the title page and the 3 pages for one candidate within the candidate PDFs. The title page shows the text and SMILES query together for organization as well as the 2D structure for the SMILES query. This ensures that chemists do not spend a lot of mental effort in trying to recognize the query from memory and instead focus on judging the actual candidates as their main focus. The main part of this exercise was to ensure each candidate got a fair assessment given a specific multi-modal query along with the neighborhood context regardless of which gene sub-type or PDF each candidate belonged to.

5.2.3 Defining the Relevance Criteria

The relevance criteria determined how relevant a candidate was when presented to the expert assessors and use those graded assessments downstream for empirically evaluating our retrieval modeling work. Our major objective was to have a graded measure for evaluation. For any kind of scientific information search, it is not expected that a single candidate is likely going to satisfy an entire query intent, specially if the intent is to find pieces of information that are loosely connected. This is also equally applicable to chemical information search where the intent can be to find various different synthesis pathways or different variations of a reaction type along with chemical or physical properties of a reactant or substrate. Such diverse all-encompassing information needs are unlikely to be met within a single passage context or even within the passage's neighborhood context. For this reason, in consultation with our group of assessors, we created a draft criteria and fine-tuned through iterations, the applicability of the criteria to the different variations of candidates we saw through the pooling process. The major facets of the iteration process was to ensure the levels included cases such as (1) A candidate containing either/both a text passage and a diagram hit, (2) Neighborhood context was being taken into account and (3) semi-relevant candidates were scored appropriately.

Table 5.2 shows the full relevance criteria definition that was adopted to score the candidate lists. One of the major contributions of this work is defining a neighborhood-based relevance level for candidates. As shown in Figure 5.4, we also show assessors the adjacent (+1/-1) pages from where the candidate passage or diagram was found. This enabled two things – (1) ensure the candidate

Table 5.2: Final Relevance Criteria for Assessing Chemical Passages and/or Diagrams for creating the graded test collection

Level	Criteria
3 **(Fully Relevant)**	Passage/Diagram **fully answers** the query for the intended information need without the need to look at other sources. *(**Fully satisfies the information need/answers the question.)*
2 **(Partially Relevant)**	Passage/Diagram **partially answers** the information need. *E.g., Passage only answers part of a synthesis pathway for a molecule, or only certain chemical or physical properties or reaction conditions. E.g., Diagram is not correct but can be used for the synthesis of an intended product.*
1 **(Somewhat Relevant)**	**Tangentially related** to the information need, but a user can look in nearby pages (±1 page) of the passage in the PDF to get their answer. If the passage/diagram is found, it can satisfy either a full/partial information need. *E.g., User is looking for the pharmacokinetic properties of a molecule but the passage describes a part of the reaction scheme for the molecule or any of the intermediate products.*
0 **(Not Relevant)**	**Not relevant at all.** *E.g., Describes a completely different reaction or molecule property.*

passage or diagram itself can be graded based on nearby context and (2) if the candidate box on Page 0 is deemed to be irrelevant, the candidate can still be relevant if any of the adjacent information contained within the adjacent pages (and any unboxed content) answers either fully or partially the intended information need of the specific query.

Set on the pooling process, the assessment process as well as the relevant criteria categories the candidates would be scored, on the next step was the multi-modal query creation and the actual process of scoring.

5.2.4 Query Creation and Scoring Process

The query creation and assessment process was done through a series of iterations. It involved an initial period of training both the expert chemist groups to mitigating personal biases and follow the criteria set. This process started with the creation of draft multi-modal queries. There was an

initial ramp up phase where the chemists did not understand the value of the SMILES part of the query because of both either unfamiliarity with SMILES representation of molecules and the need for an additional part of a query (in lieu of the textual part) because of unfamiliarity of preparing multi-modal queries in their usual process of literature search using Reaxys or SciFinder.

The draft queries submitted by the experts first were either too complex or too simple for our pooling process. Many of the queries designed initially were based on looking at a specific sub-type of PDFs (for e.g., only looking at PDFs that catered to the GLP-1 gene). To mitigate this bias, the chemists were directed to briefly go through the entire collection of PDFs that were curated and figure out similar patterns and reaction schemes. This was an essential part of the collection process as both parties realized that the sort of information need intent they formed were based on incomplete context as the collection was comprised of pharmaceutical patents targeting gene inhibitors and many of them included similar reaction schemas as well as similar reaction types and surprisingly, many closely related molecule structures as well. This went a long way in mitigating the implicit bias in our assessors and delink the query creation process context and the final assessment candidate list context as candidates that appeared from a different gene than originally intended were also more likely to be objectively judged based only on the relevance criteria defined above.

Another issue with the draft queries that were created was the wording of the textual part of the query. Some of the queries formed were biased because of the document formatting of relevant information. For e.g., some reactants, products and substrates were referred to by their local schema numbering such as “substrate 2a” or “reagent 5c”. These were later corrected by replacing these with the actual chemical name, IUPAC name, common name or abbreviations. For e.g., “reagent 5c” was replaced by the IUPAC name Ethyl Ethanoate or the common name Ethyl Acetate or the abbreviation ETOAc.

After several rounds of trial and error, which involved getting the queries from the experts, running the pooling process and asking the asking the chemists to judge the candidates, the training process was completed. For the final query assessment phase, the expert group from RIT comprising of one professor and one senior undergrad student were tasked with creating and assessing candidates for 10 different queries with each person scoring 5 queries each. The group from UR comprising of two professors and two senior PhD students were tasked with creating and scoring 25 queries.

Table 5.3: Test Collection Statistics for Multi-Modal Relevance Assessment.

Metric	Value
Total Multi-Modal Queries Created	35
Total Candidates Assessed	655
Number of Words (Mean)	7.4
SMILES Length (Mean)	30.4
% Fully Relevant Candidates (Score 3)	5.6%
% Partially Relevant Candidates (Score 2)	17.7%
% Tangentially Relevant Candidates (Score 1)[1]	20.8%
% Not Relevant Candidates (Score 0)	55.9%

[1] Rating of 1 means a hit of score 2/3 within the 3 returned pages.

5.2.5 Final Test Collection Statistics

A total of **35** queries were created and annotated by the assessors. **Appendix D** shows the exhaustive list of all multi-modal queries created along with their candidate type breakdown (Text-Only, Diagram-Only, Text+Diagram). This major collection ensured that a graded relevance test collection could be used for evaluating different models pertaining to chemical literature search at the passage, diagram and page level granularity. The topics, wording of the queries as well the type of scores given to candidates had plenty of diversity in them. Topics were chosen to be either asking for a specific type of reaction such as "Buchwald-Hartwig coupling with piperidine" or asking for a specific reagent and reactant combination such as "Show me where this reagent, 1-ethyl-3-(3-dimethylaminopropyl)carbodiimide(ethyldimethylaminopropylcarbodiimide), is utilized to couple an amine with carboxylic acid". This ensured that the initial goal of the retrieval approach being used to cater to varying style of information need could still be supported and evaluated.

Table 5.3 shows the final statistics of the created test collection. We see the effect of the initial training iterations with both the expert groups come to good effect where the quality of the final collection is sound. A good balance between relevant (>= 1) candidates being 45% and not-relevant candidates suggests that both the assessor training as well as the pooling mechanism were well aligned with the task and type of information needs. Our model and documents could faithfully represent the information intent in the queries. This shows that the collected assessment data is a real usable graded relevance test collection that can be used to evaluate retrieval models not only

for this research but creates a benchmark for the scientific community to use and evaluate their own modeling approaches.

5.3 Summary

This chapter described the data extraction improvements that led to more coverage of passages throughout the raw PDF collection and was augmented by a secondary external knowledge base curated from PubChem that in addition added to the number of SMILES extracted within the raw passages. These improvements led to the development of a large index which could be used to pool candidates across the PDFs with candidates that were generally considered to relevant for chemist designed queries. A new method of judging relevance by assessing +/- 1 pages in context from a candidate hit page provided a new way of looking at retrieval candidates for document based retrieval.

In the next chapter, we introduce our modeling and optimization approach to evaluate our test collection. We introduce a new method to train with unpaired text passages that can be trained in union with paired samples together to enable our model to learn more chemical context when a lot of the data does not contain annotated pairings between the text and the SMILES modality. The experiments done demonstrate the effectiveness of our approach towards training with pseudo-positive pairings that do not completely break the embedding space and perform as well as training with only paired data.

Chapter 6

Semi-Supervised Contrastive Learning with Pseudo-Positive Text-Molecule Pairs

This chapter presents experiments addressing RQ2 and RQ3 as stated in Section 1.2.2. RQ2 asks how we can address the cross-modal alignment between textual and SMILES representations while RQ3 asks about methodologies that can be used to train when some text-SMILES pairs contain missing data. To address RQ2, we train a cross-modal contrastive model aligning chemical text passages with SMILES representations and evaluate retrieval quality on our expert-annotated graded relevance test collection across a range of temperature and architectural configurations. The investigation of RQ3 is motivated directly by an empirical observation from our passage extraction and SMILES conversion pipeline: a substantial fraction of extracted passages failed to yield any valid SMILES pairing, arising not from extraction failures but from the absence of resolvable chemical entity mentions in contextual patent text. Rather than discarding these unpaired passages, we treat their prevalence as a research opportunity — can a contrastive model leverage chemically contextual but molecularly ungrounded text (without valid positive SMILES correspondence) to learn richer representations, while preserving the embedding space alignment established by true passage–SMILES pairs? We evaluate this question by systematically characterizing the conditions — in terms of missing sample ratio, pseudo-positive selection strategy, loss weighting α, and temperature τ — under which auxiliary context from unpaired passages helps, is neutral, or actively harms retrieval performance as measured by R'@k and nDCG'@k on our graded relevance collection.

6.1 InfoNCE Loss and Missing Modality Hard-Negative Alignment

We now introduce the semi-supervised missing modality loss that is used to align valid positive pairs from both modalities with added weak supervision from contextual text passages without a SMILES pairing. We also introduce the modeling approach that was taken to create the valid pairings. The prelude to that is a description of existing approaches in scientific literature that caters to this problem of missing valid pairings in a cross-modal alignment paradigm and how our approach is fundamentally different and tackles a different problem of completely unsupervised textual context data.

6.1.1 Background

We plan to tackle RQ3 (missing modality alignment) by generating a variation of the InfoNCE [142] contrastive loss. InfoNCE stands for Information-Noise Contrastive Estimation and has been recently used in training joint multi-modal representation learning. Formally, given a positive pair of samples, it computes the ratio of the signal-to-noise ratio between the similarity of the positive pair compared to a batch of negative pairs. A positive pair in this context is a text passage and its corresponding molecule named entity as the SMILES representation. E.g., "benzene is a flammable compound" with "c1ccccc1". The similarity is usually computed as the cosine distance between the vector embeddings in the pairs. For more details about this loss, please refer to Chapter 2 Section 2.4.4. Given two modalities A and B, a mini-batch of N samples, the InfoNCE loss for modality A compared to modality B becomes:

$$\hat{L}_{A\to B} = -\frac{1}{N}\sum_{i=1}^{N} log\frac{exp(sim(z_i^{A+}, z_i^{B+})/\tau)}{exp(sim(z_i^{A+}, z_i^{B+})/\tau) + \sum_{j=1}^{N-1} exp(sim(z_i^{A+}, z_j^{B-})/\tau)} \tag{6.1}$$

z_i^{A+}, z_i^{B+} is the positive pair where z_i^{A+} is the i^{th} sample of the mini-batch belonging to modality A and z_i^{B+} is the corresponding positive sample from modality B. Thus, the numerator forms the signal part of the loss. For each positive sample from A, all the other examples from B in the mini-batch become the negative examples and form the noise part of the loss. As shown in Equation 6.1, the similarities between the i^{th} sample from modality A: z_i^{A+} and the other negative samples from modality B, z_j^{B-} are summed up and form the noise. In simpler terms, this means that we want to make sure that the positive pair in latent space is as far apart as possible from the negative pairs

that are made from each positive sample from modality A and the other negatives from modality B in the mini-batch. τ or the temperature parameter is used as a normalization constant and is commonly set as a constant and tuned or learned during training. Most works using this loss have set τ between [0.01, 0.2] depending on the task, dataset and model. Most prominent works like SimCLR [15, 14] and CLIP [117] set it to 0.07. Additionally, it acts as a weighting parameter to focus on harder negatives or treat all negative pairs equally (discussed in the next paragraph). Thus τ is used as parameter to weigh hard-negatives more or less.

$sim(z^A, z^B)$ is the cosine similarity or dot product of the vector embeddings from the modalities A and B. They are unbounded and range from $[-\infty, \infty]$. This in practice does not make a difference as the nature of the InfoNCE loss is a softmax operation where the positive pair signal (numerator) is divided against the sum of all the losses (positive pair + all the negative pairs) (denominator). This is exactly the softmax operation. Thus the total InfoNCE loss mathematically is always bounded between [0,1]. The τ operator is a normalization parameter that can be varied to make the softmax sharper (focusing more on hard negatives) or softer (treating all negatives equally). Mathematically, if τ is small, it makes the term $exp(z/\tau)$ larger. Conversely, if τ is large, it makes the terms softer. Therefore, in the denominator of the InfoNCE loss in Equation 6.1, for the cases where the similarity $sim(z^{A+}, z^{B-})$ is larger for a negative pair (hard-negative), a smaller τ will increase the total loss contribution from these pairs.

It is important to note that InfoNCE loss for two modalities A and B are computed twice, comparing each positive sample to every negative sample in the other modality (the denominator changes). Therefore the total loss becomes the average of $\hat{L}_{A\to B} + \hat{L}_{B\to A}$. Therefore, the InfoNCE loss measures the signal-to-noise ratio, and the training objective aims to maximize this objective. This is the reason there is a negative summation as shown in Equation 6.1.

6.1.2 Previous Work On Training With Missing Modalities

Previous works have addressed missing modality in different ways. Qiu et al. [116] and Ma et al. [87] used masked or dummy inputs for a missing modality and ensuring that the available training samples help in aligning the complete pairs. Lee et al. [69] only assumed that modalities might be missing only at inference time but the complete pairs were available during training time for multi-modal retrieval. Other works like [156], [152], [76], [75], [42], [128] approached the problem from the perspective of the objective function. Some completely ignored the samples with missing samples from consideration while other considered a weighing sample for pairs with weak or incorrect positives. We will discuss the differences between the approaches in this section and

discuss the major differences in our approach in Section 6.1.3.

Chuang et al. [19] first tried to mitigate the effect of randomly sampling positive examples as negative examples because there might be inadvertent addition of another positive example in the negative samples due to datasets not being fully labeled. E.g., for a positive pair containing the text 'cat' and an image of a cat, when randomly sampling negative images for training, there might be other cat images that have not been labelled as a cat image. They sampled the negatives as well as likely positives that mitigated the effect of unstable training through weakly paired positive and negative samples. Wu et al. [151] considered each sample as its own class, and the objective was to minimize each sample to every other sample and maximize similarity with itself. MMP [104] had a similar constraint to our problem where only a specific modality had a problem of missing positive samples instead of it being generalized to any modality. To address this, they used a conditional projection head and loss along with the encoders for each modality. For the pairs that were missing the positive sample for a pair, they did not use the projection head and just used a masked input for the missing modality and treated the mask as a pseudo-positive pair. For pairs that had the missing modality, the projection head was used 30% of the times. In those cases, the model was forced to learn what the input was supposed to be from the other two modalities. This was done by using a masked input again for the encoder missing a sample but using the projection head to predict the actual input. An L2 similarity loss was used between the predicted output of the projection head and the ground-truth sample. This idea was to align the encoder that sees the masks to interpret what the likely input is supposed to be from understanding the other two modalities.

6.1.3 Weak Supervision Approaches vs. Our Approach

Compared to our approach of hard-negative based supervision, the prior works assumed a much more constrained set of experiment parameters in their work. Prior work on cross-modal contrastive learning between chemical text and molecular representations has largely assumed the availability of clean, curated positive pairs. MoleculeSTM [78] constructs over 280,000 structure–text pairs from PubChem, where each text description is explicitly linked to a canonical molecular structure, and trains a joint embedding space using a standard InfoNCE objective over these well-defined pairings. Similarly, ACML [146] and CLAPS [144] operate under the assumption that positive supervision is either directly available or can be derived from structural similarity metrics applied to fully annotated molecular datasets. In the broader contrastive learning literature, methods addressing imperfect supervision — such as puCL [1] and the framework of Cui et al. [26] — handle weak

supervision at the *label* level, where class memberships are noisy or partially observed, but still assume that each training sample has a well-defined candidate positive from which to construct the contrastive objective. Hard negative mining approaches in cross-modal retrieval, including SAHN [74] and CrossCLR [166], focus on improving the quality of the *negative* set by identifying and either reweighting or removing false negatives, but again presuppose that positive pairings themselves are reliable and complete.

Our Approach: Our work addresses a fundamentally different and practically important scenario that none of these approaches handles: a large-scale scientific corpus in which a substantial fraction of passages are chemically relevant — drawn from drug patents and publications — but contain no extractable molecular entity mentions, and therefore have no ground-truth positive SMILES pairing. Rather than discarding these passages, we propose treating the hardest in-batch negative SMILES as a pseudo-positive supervision signal for unpaired passages, incorporating this signal through an auxiliary weighted InfoNCE objective controlled by a loss weighting parameter α. This formulation is distinct from PU learning [1] in that we do not assume any prior over which SMILES might be relevant to an unpaired passage — the pseudo-positive is selected purely through geometric hard negative mining in the current embedding space at training time, without any external label information. It is also distinct from false negative mitigation approaches [74, 166] in that we are not correcting a noisy negative set but rather constructing a positive signal where none previously existed. Crucially, our empirical results demonstrate that the *choice* of pseudo-positive assignment is decisive: random SMILES assignment as pseudo-positive degrades retrieval performance relative to training on valid pairs alone, while hard negative selection consistently improves ranking quality as measured by nDCG@10, confirming that geometric proximity in the learned embedding space provides a meaningful pseudo-supervision signal for contextual chemical passages even in the absence of explicit molecular entity mentions. To our knowledge, this is the first work to systematically study pseudo-positive contrastive supervision for partially unpaired cross-modal chemical retrieval, and to empirically characterize the role of loss weighting, missing sample ratio, and temperature in governing whether such pseudo-supervision helps or hurts downstream retrieval quality.

6.1.4 Hard-Negative Alignment with InfoNCE

RQ3 as stated in Section 1.2.2 is to understand whether the addition of additional samples from one modality with no absolute positive corresponding sample in the other modality has a positive

or negative impact on the final model performance.

To examine this, we added additional passages to our training dataset where the valid pairs were used along with pairs where the text modality did not contain a corresponding positive SMILES. To ensure that this training paradigm worked, we split the InfoNCE loss into two components as shown by the equations below.

$$\hat{L}^{\mathbf{Pos}}_{A\to B} = -\frac{1}{P}\sum_{i=1}^{P}\log\frac{\exp(sim(z_i^{A+}, z_i^{B+})/\tau)}{\exp(sim(z_i^{A+}, z_i^{B+})/\tau) + \sum_{j=1}^{P-1}\exp(sim(z_i^{A+}, z_j^{B-})/\tau)}$$
$$\text{where } z_i^{B+} \text{ is the ground-truth positive SMILES for passage } i$$
$$\text{applied only to the } P \text{ passages with valid ground-truth SMILES pairings} \tag{6.2}$$

$$\hat{L}^{\mathbf{Neg}}_{A\to B} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(sim(z_i^{A+}, z_i^{\mathbf{B*}})/\tau)}{\exp(sim(z_i^{A+}, z_i^{\mathbf{B*}})/\tau) + \sum_{j=1}^{N-1}\exp(sim(z_i^{A+}, z_j^{B-})/\tau)}$$
$$\text{where } z_i^{\mathbf{B*}} = \arg\max_{j\neq i} sim(z_i^{A+}, z_j^{B-}) \text{ is the hardest negative SMILES assigned as pseudo-positive}$$
$$\text{applied only to the } N \text{ ungrounded passages with no valid SMILES pairing} \tag{6.3}$$

P is the number of positive samples in the batch and N is the number of negative samples in the batch. z_i^{B*} becomes the hardest negative SMILES sample found for a text passage within the batch.

The total loss becomes:

$$\hat{L}^{\mathbf{Total}}_{A\to B} = \hat{L}^{\mathbf{Pos}}_{A\to B} + \alpha\hat{L}^{\mathbf{Neg}}_{A\to B} \tag{6.4}$$

For SMILES to text direction, we simply only use the reverse the modalities in Equation 6.2 and ignore the pseudo positive loss. That means only the valid SMILES in the batch are compared against every text sample to create the denominator. Equation 6.3 is ignored in this direction as there does not exist any valid SMILES that do not have a positive text pairing in our setup (no pseudo positive text passages are needed).

The loss here is the sum of the positive and hard-negative InfoNCE losses. We realized that adding the hard negative samples as pseudo positive poses a challenge due to resulting noise arising from

the samples. To alleviate this convergence issue, we use an extra α parameter to control the impact of the loss from the hard-negative InfoNCE towards the total loss. The balance between the loss contributions is decided by the distribution of full pairs and missing pairs per batch. The balance is maintained by keeping the distribution of samples consistent across batches. This means if the number of missing samples is set to 50%, half of each batch will be valid pairs and the other half will be missing pairs. In our experiments, we try with {10%, 20% and 50%} missing samples per batch.

We do not go higher with the number of missing samples because the hardest negative SMILES pairing for the missing sample pairs are obtained from the valid SMILES pairings. If the ratio of valid pairings to the total becomes too small, the hard negatives chosen for each of the text samples without pairings will start to point to the same valid SMILES. This inherently will add more noise to the overall system as the SMILES chosen will not really be "hard" and will mis-align the embedding space.

The α parameter is further used to control the effect of the noisy psuedo-positive samples. This is because, at training initiation time, the hard-negatives chosen in terms of cosine similarity of the embeddings are completely random. The only useful signal comes from the valid pairs positive component of the loss. If the negative component is not tempered, the embedding alignment breaks down and the model fails to converge well reinforcing noisy random hard negatives as the true-positive. Thus applying this alpha parameter helps to stabilize the training over time.

6.1.5 Modeling Approach

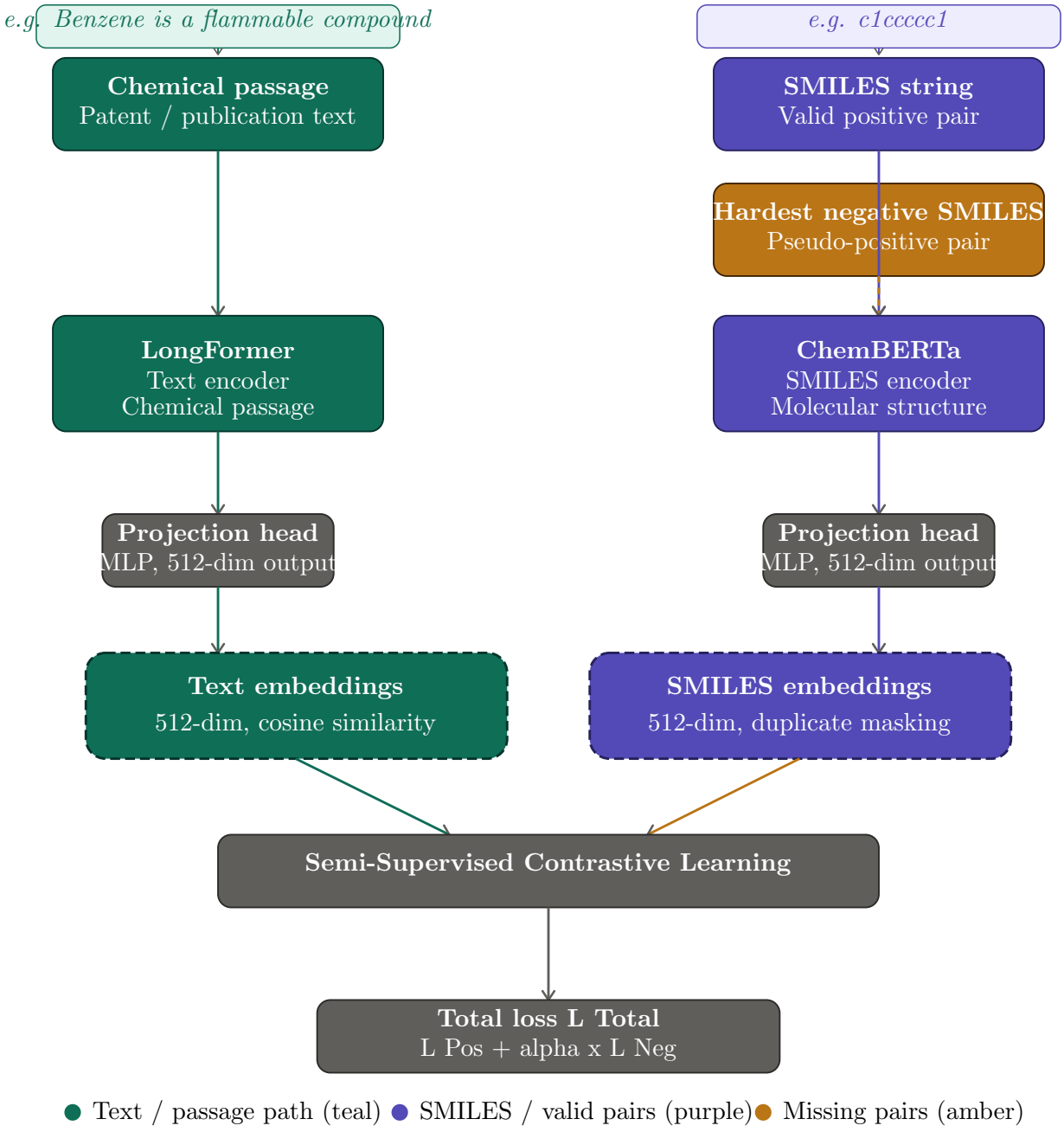

Figure 6.1: Modeling Architecture: Chemical passages are encoded via LongFormer and SMILES strings via ChemBERTa, both projected to a shared 512-dimensional embedding space. Valid passage–SMILES pairs (**336K**) contribute a standard InfoNCE loss $\hat{L}^{\text{Pos}}$, while passages lacking a ground-truth SMILES pairing (**44k**) are assigned the hardest in-batch negative SMILES as a pseudo-positive, contributing $\alpha\hat{L}^{\text{Neg}}$. The total loss is $\hat{L}^{\text{Total}}_{A\to B} = (\hat{L}^{\text{Pos}}_{A\to B} + \alpha\hat{L}^{\text{Neg}}_{A\to B})$.

Figure 6.1 shows our overall modeling setup. In a CLIP-style way, we created an aligned shared representation between chemical passage texts using the LongFormer [8] and ChemBERTa [18] for encoding text and SMILES.

- **SMILES encoder:** ChemBERTa, [18], will be used as the molecule encoder in our joint model. It has the same architecture, input and output dimensions as BERT [30] and originally was used for downstream molecular property classification and prediction tasks from a SMILES input. It was trained on 10M unique SMILES obtained from PubChem and therefore already has foundation knowledge about latent space semantic properties of molecules. GraphT5 [59] is the latest model to use SMILES as an input but they used an additional modality of molecular graphs as an input and unlike ChemBERTa, trained on only 324k unique SMILES from the PubChem-300k [163] dataset. Another popular model MolTRES [108] used a generator-discriminator architecture and was not suitable for our joint training task. Other models used for chemical modeling represented the molecule input as a 2D graph which was again not applicable to our approach of using only SMILES as an input.

- **Chemical Passage Encoder:** SciBERT [7] would be used as the passage encoder in our joint model. Compared to BERT, it also has the same input, architecture and output space but the major differences being the token vocabulary and pretraining dataset. SciBERT was trained on 1.14M domain biochemical and computer science based domain specific papers from Semantic Scholar. Additionally, unlike BERT which was trained on a general token vocabulary set, SciBERT was trained on a scientific specific vocabulary set but the total size of both was around 30k tokens. However, both used the WordPiece tokenizer [134] with the difference being the original word set that was used to develop the vocabulary. Concretely, a word `phosphorylation` would be tokenized as `['ph', 'os', 'ph', 'ory', 'lation']` using the BERT tokenizer but SciVocab would consider the entire word `['phosphorylation']` as a single token.

  We quickly realized early on that the fixed input token length supported by SciBERT (512 token limit) was not enough to contain the embeddings of some of the long context passages within our test collection. To mitigate this, we migrated the SciBERT weights to a longer context BERT-styled model called LongFormer [8] to ensure that we still preserve implicit chemical knowledge contained to help speed up convergence when training with ChemBERTa jointly.

6.2 Experiments

We now introduce the experiments that we performed on our training and expert annotated collection.

Dataset: Our training corpus comprises 373K valid positive passage–SMILES pairs and 44K contextual passages with no extractable chemical entity mentions and therefore no ground-truth SMILES pairing. Of the 373K valid pairs, 336K were used for training and the remaining 37K (approximately 10%) were held out for validation. The 44K unpaired passages were incorporated into training in varying proportions to evaluate our pseudo-positive supervision hypothesis. Both the training and validation splits exhibit substantial structural complexity: a single SMILES string is on average paired with at least four distinct chemical passages, reflecting the many-to-one nature of chemical entity mentions across patent text. Concretely, of the 37K validation pairs, only approximately 8K SMILES are unique, making the retrieval task considerably more challenging than a standard one-to-one matching benchmark and closely mirroring the redundancy found in real-world document collections. We experiment with varying ratios of valid and unpaired samples at training time to characterize the effect of pseudo-positive supervision on cross-modal and unimodal retrieval performance.

Experimental Setup & Hardware: Majority of our experiments were run on the National Center for Supercomputing (NCSA) cluster system. Some specifics of the hardware used for training each experimental run are given below.

1. CPU: Intel Xeon Platinum 8558
2. No. of Cores Used: 8
3. CPU Memory Used: 256GB
4. GPU: NVidia H200 (141GB VRAM)
5. Storage (Shared): 1TB

Each of the training and validation experiments were run using the same hardware resources to maintain an apples-to-apples comparison. To ensure easy management of model runs and hyperparameter variations, we used Weights and Biases to keep track of our experimental runs and model

checkpoint files. All of runs used a constant batch size of 50 to maintain parity in our scores as self-supervised training using InfoNCE loss is extremely sensitive to the total number of examples seen in each mini-batch to ensure final scores are only based on other independent variables apart from the batch size used. We train each iteration for 20 epochs. As shown in Table 6.1, we use the held out validation set containing 37K examples as the test set for training-time evaluation and various ratios of positive and missing pairs for training.

All of our experiments were trained from the initial pre-trained weights of SciBERT and ChemBERTa and each run was trained independently of each other jointly from scratch to ensure there was no impact of different training paradigms in our experiments. Specifically, the effect of additional training samples used were made to not interfere with the effect of changing pseudo-positive weighting or the effect of the temperature parameter τ. For retrieval we create separate embedding tables for text and SMILES modalities, with the idea that each modality embedding space has joint context from both modalities.

Briefly, these are the different independent variables we experiment with, tuning one and keeping the others constant.

1. Number of Positive Pairs (Total) – [336k, 302k, 275k, 172k]
2. Number of Ungrounded Text Passages – [0, 44k]
3. Positive-Pseudo Positive Sample Ratio – [0%, 10%, 20%, 50%]
4. Temperature (τ) – [0.1, 0.15, 0.2]
5. Pseudo-Positive Loss Weighting (α) – [0.1, 0.5, 0.8, 1.0, 2.0]

Evaluation metrics: All experiments are evaluated against the same graded relevance test collection of 35 expert-annotated queries described in Section 5.2.2. Since each query has between 10 and 30 annotated candidates, we report nDCG'@1, 5, 10 and R'@1, 5, 10 as our primary retrieval metrics. Recall at rank k (R@k) measures the proportion of relevant documents successfully retrieved within the top k results, reflecting the breadth of the retrieval system's coverage at a given cutoff. Normalized Discounted Cumulative Gain at rank k (nDCG@k) is a ranking quality metric that additionally accounts for the position of relevant documents within the top k results, assigning higher scores when more relevant documents appear closer to the top of the ranked list and incorporating graded relevance judgments where available.

Table 6.1: Training data composition across experimental conditions. Missing % denotes the proportion of unpaired passages sampled per batch.

Condition	**Available Positive pairs (Text → SMILES)**	**Available Ungrounded text-only passages**
0% missing (control)	336K	0
10% missing	302K	44K
20% missing	275K	44K
50% missing	172K	44K
Validation	37K	0

The prime notation (R'@k and nDCG'@k) denotes that both metrics are computed against an incomplete relevance judgment set — that is, only the subset of retrieved candidates that were explicitly assessed by human annotators are considered when computing the metric, rather than assuming all not considered candidates are non-relevant. This is standard practice in test collections constructed via pooling, where it is infeasible to assess every document in the corpus, and prevents the metrics from unfairly penalizing systems that retrieve unjudged but potentially relevant candidates.

For multi-modal queries, we follow the same pooling mechanism used during test collection construction, with one key difference: rather than ensembling multiple models per modality, we use a single LongFormer embedding index for text retrieval and a single ChemBERTa embedding index for SMILES retrieval. Candidates from both indices are fused using Reciprocal Rank Fusion (RRF), after which the top-1000 results are filtered to retain only those candidates present in the annotated pool. No bucketing strategy is applied at inference time, as the annotated candidate set is fixed and exhaustive for each query.

Baselines: We chose two sets of different control experiments to evaluate our methodology. The first natural baseline was chosen to be the standard run where we only train using noise-free, high fidelity samples using the entire set of 373k positive paired samples. Standard InfoNCE loss was used and pseudo-positive text passages were discarded. For our second control experiment, to evaluate the efficacy of hard-negative pseudo-positive pairing strategy, we use our additional objective function but instead of pairing with hardest negative SMILES, we pair each ungrounded text passage with another random SMILES. Our different data distributions for varying the amount positive versus missing samples are shown in Table 6.1.

Table 6.2: Impact of temperature (τ) and missing sample ratio per-batch on cross-modal retrieval performance (Text $\rightarrow$ SMILES). All experiments use a fixed α of 0.5. The control (no missing) group uses standard InfoNCE loss over valid pairs only. The random negative control replaces hard negative mining with random in-batch SMILES selection as the pseudo-positive.

Missing %	**Condition**	τ	**R'@1**	**R'@5**	**R'@10**	**nDCG'@1**	**nDCG'@5**	**nDCG'@10**
0% (no missing) 336k +ve + 0k miss	Standard InfoNCE	0.10	0.058	0.325	**0.553**	0.247	0.364	0.456
		0.15	0.077	0.318	0.495	0.308	0.378	0.445
		0.20	0.093	0.354	0.477	**0.379**	**0.422**	0.451
50% 172k +ve + 44k miss	Random negative	0.10	0.093	0.320	0.520	0.288	0.375	0.460
		0.15	0.072	0.302	0.457	0.308	0.364	0.416
		0.20	0.067	0.305	0.429	0.288	0.373	0.400
10% 302k +ve + 44k miss	Hard negative	0.10	**0.100**	**0.339**	0.533	0.308	0.398	**0.470**
		0.15	0.073	0.313	0.456	0.268	0.364	0.420
		0.20	0.071	0.298	0.454	0.298	0.358	0.423
20% 275k +ve + 44k miss	Hard negative	0.10	0.094	0.332	0.521	0.288	0.393	0.461
		0.15	–	–	–	–	–	–
		0.20	0.067	0.293	0.456	0.288	0.348	0.409
50% 172k +ve + 44k miss	Hard negative	0.10	0.055	0.305	0.508	0.197	0.349	0.429
		0.15	0.067	0.317	0.456	0.258	0.366	0.413
		0.20	0.061	0.314	0.450	0.227	0.362	0.402

6.2.1 Impact of Temperature τ

Table 6.2 presents retrieval performance across all experimental conditions, varying the per-batch missing sample ratio, pseudo-positive selection strategy, and temperature τ. It is important to note that the missing % refers to the proportion of ungrounded passages sampled *per batch* at training time, not the overall ratio of ungrounded to valid passages in the dataset. The full pool of 44K ungrounded passages is available across all non-zero missing conditions; what varies is how many of these passages are drawn into each batch relative to valid passage–SMILES pairs. A 10% missing ratio therefore means that each training batch is composed of 90% valid pairs and 10% ungrounded passages, not that only 10% of the ungrounded passages are used overall. We discuss the results with respect to three axes of comparison: the effect of hard negative pseudo-positive supervision against both control baselines, the role of temperature, and the impact of the per-batch missing sample ratio.

Hard negative supervision vs. control baselines. Before comparing retrieval metrics directly, it is important to acknowledge a fundamental asymmetry in the training data available to each experimental condition, as shown in Table 6.1. The standard InfoNCE control (0% missing) trains exclusively on valid passage–SMILES pairs and has access to the full 336K high-fidelity paired samples per epoch. By contrast, the 10% missing hard negative condition draws from 302K valid pairs and supplements each batch with ungrounded passages, meaning it sees fewer ground-truth paired samples overall. This distinction is critical when interpreting the comparative results: the standard InfoNCE control benefits from a larger and cleaner supervision signal, and any competitive or superior performance by the pseudo-positive conditions must therefore be understood as being achieved with strictly less high-fidelity paired data. That the hard negative conditions remain competitive under this disadvantage is itself a positive finding.

With this context in mind, the 10% per-batch missing condition with hard negative pseudo-positive supervision at $\tau = 0.10$ achieves R'@10= 0.533 and nDCG'@10= 0.470. The standard InfoNCE control at the same temperature achieves a higher R'@10 of 0.553 but a lower nDCG'@10 of 0.456, despite having access to 34K more high-fidelity paired samples. This difference between recall and ranking quality reflects a meaningful qualitative difference in what the two objectives learn: the standard InfoNCE loss, trained on a larger set of valid pairs, maximises broad coverage of the candidate pool, while the hard negative pseudo-positive objective encourages the model to rank relevant SMILES more confidently near the top of the retrieved list even with a reduced valid pair budget. The improvement in nDCG'@10 under the 10% hard negative condition is therefore particularly notable, as it is achieved against a baseline that has a data volume advantage, and suggests that the pseudo-positive objective contributes genuine representational enrichment beyond what additional valid pairs alone provide.

A particularly informative comparison is between the 50% hard negative and 50% random negative conditions, as both use identical per-batch missing ratios and differ only in their pseudo-positive selection strategy. The random negative control, which assigns a randomly sampled in-batch SMILES as the pseudo-positive, achieves a surprisingly competitive R'@10= 0.520 and nDCG'@10= 0.460 at $\tau = 0.10$, approaching the standard InfoNCE baseline despite using only 172K valid pairs per epoch — nearly half the data budget of the control. This result warrants careful consideration. We hypothesise that the random negative pseudo-positive loss, rather than providing a meaningful alignment signal, acts primarily as a *regularizer* on the contrastive objective: by randomly assigning chemically unrelated SMILES as pseudo-positives for ungrounded passages, the loss introduces a consistent but non-directional gradient perturbation that prevents the embedding space from overfitting to the most frequently occurring valid pair samples. Unlike hard negative selection, where

the pseudo-positive is specifically chosen to be maximally confusable with the text embedding, the random negative introduces no systematic bias in any chemical direction, and its gradient contribution is easily overridden by the dominant valid pair signal. This regularization effect appears to be beneficial at the 50% missing ratio where the valid pair budget is most reduced, partially compensating for the smaller set of high-fidelity pairs seen per epoch. The 50% hard negative condition, by contrast, achieves R'@10= 0.508 and nDCG'@10= 0.429, falling below the random negative control on both metrics. At this missing ratio, the hard negative pseudo-positive — chosen for its geometric proximity to the text embedding — introduces directional gradient signal that conflicts with the valid pair objective at too large a scale, degrading performance in a way that the non-directional random negative does not. Critically, both 50% conditions underperform the standard InfoNCE baseline on R'@10, confirming that neither pseudo-positive strategy is sufficient to compensate for the reduced valid pair budget at high missing ratios, and that the proportion of ungrounded passages sampled per batch is the dominant factor governing performance at high ratios.

Effect of temperature. Across all per-batch missing ratios and both pseudo-positive strategies, $\tau = 0.10$ consistently produces the strongest retrieval performance, with R'@k and nDCG'@k degrading monotonically as τ increases to 0.15 and 0.20. This pattern is consistent with the known effect of temperature in InfoNCE-based contrastive learning: a lower temperature sharpens the softmax distribution over negatives, forcing the model to be more discriminative and producing a better-separated embedding space that supports precise ranking. Higher temperatures produce a flatter gradient signal that is insufficient to resolve the fine-grained structural similarity required for chemical retrieval, particularly given the high degree of SMILES redundancy in the training corpus.

Effect of per-batch missing sample ratio. Performance under hard negative supervision degrades progressively as the per-batch missing ratio increases beyond 10%, with the 50% condition recording the lowest scores across all temperatures. This degradation is compounded by the reduction in valid pair budget: as shown in Table 6.1, the 50% missing condition trains on only 172K valid pairs compared to 302K for the 10% condition and 336K for the control, meaning the model receives both a weaker pseudo-positive signal and less high-fidelity paired supervision simultaneously. At a per-batch missing ratio of 50%, half of the gradient signal at every update step originates from pseudo-positive alignments rather than ground-truth pairs, introducing noise that outweighs the representational benefit of exposing the model to additional ungrounded passages.

Table 6.3: Impact of the pseudo-positive loss weight α on cross-modal retrieval performance (Text $\rightarrow$ SMILES) at a fixed $\tau = 0.20$ and 50% per-batch missing ratio. The control uses random in-batch SMILES as the pseudo-positive with $\alpha = 1.0$.

Condition	$\boldsymbol{\alpha}$	**R'@1**	**R'@5**	**R'@10**	**nDCG'@1**	**nDCG'@5**	**nDCG'@10**
Random negative (control)	1.0	0.093	0.320	0.520	0.288	0.375	0.460
Hard negative	0.1	0.065	0.294	0.503	0.247	0.353	0.434
	0.5	0.088	0.324	0.528	0.258	0.369	0.451
	0.8	0.093	0.326	0.530	0.288	0.378	0.460
	1.0	**0.093**	0.324	**0.547**	**0.288**	**0.380**	**0.467**
	2.0	0.091	**0.327**	0.534	0.247	0.371	0.460

The 20% per-batch condition occupies an intermediate position, performing comparably to the 10% condition at $\tau = 0.10$ on R'@10 but falling short on nDCG'@10, suggesting that a moderate per-batch missing ratio preserves the embedding space alignment established by valid pairs while still benefiting from supplementary pseudo-positive supervision. Taken together, these results indicate that hard negative pseudo-positive supervision is most effective when the per-batch missing ratio is kept moderate ($\leq 20\%$) and the temperature is maintained at $\tau = 0.10$, producing a training regime in which the contrastive objective remains anchored by high-fidelity valid pair supervision while gaining representational coverage from contextual ungrounded passages.

6.2.2 Impact of alpha

Table 6.3 presents the effect of the pseudo-positive loss weight α on retrieval performance, holding the per-batch missing ratio fixed at 50% and $\tau = 0.20$. These conditions represent the most challenging setting examined in this work, with half of every training batch composed of pseudo-positive alignments and a softer temperature providing less discriminative gradient pressure. The ablation isolates the role of α in governing the relative contribution of the pseudo-positive loss term to the total training objective.

Effect of α on hard negative supervision. At low pseudo-positive weight ($\alpha = 0.1$), the hard negative condition underperforms the random negative control across most metrics, recording R'@10= 0.503 and nDCG'@10= 0.434 against the control's R'@10= 0.520 and nDCG'@10= 0.460. This is consistent with the finding from Table 6.2 that at high per-batch missing ratios, a low pseudo-

positive weight is insufficient to provide a useful alignment signal and instead introduces a small but persistent gradient perturbation that marginally disrupts the valid pair objective. However, as α increases, the hard negative condition improves monotonically across all metrics, overtaking the random negative control at $\alpha = 0.5$ on R'@10 (0.528 vs. 0.520) and matching it on nDCG'@10, before surpassing it convincingly at $\alpha = 1.0$ with R'@10= 0.547 and nDCG'@10= 0.467. This monotonic improvement with increasing α demonstrates that the hard negative pseudo-positive signal becomes progressively more beneficial as its contribution to the total loss is amplified and the peak performance lies somewhere between 1.0 and 2.0. and that the lower performance of hard negative supervision at 50% missing observed in Table 6.2 was partly an artifact of insufficient loss weighting. The other obvious part was because it saw a lot less positive paired samples than other experiments to better anchor the embedding space.

Hard negative vs. random negative control. The contrast between the two conditions at $\alpha = 1.0$ is particularly revealing. At equal weighting, the hard negative condition outperforms the random negative control on most metrics, with gains of +0.027 on R'@10 and +0.007 on nDCG'@10. This result reinstates the claim that geometric hard negative mining provides a stronger pseudo-positive supervision signal than random SMILES selection, while also demonstrating the conditions under which this advantage is most visible: the hard negative signal requires sufficient loss weight to overcome the noise introduced by the high per-batch missing ratio before its selectivity advantage over random assignment becomes apparent. Taken together with the findings from Table 6.2, these results suggest that the interaction between α and the per-batch missing ratio is the critical design consideration for pseudo-positive contrastive supervision: hard negative mining outperforms random selection when either the missing ratio is kept moderate or the loss weight is sufficiently high to ensure the pseudo-positive gradient dominates over noise, but degrades relative to random selection in the low-weight, high-missing-ratio regime where the signal is too weak to be directional.

6.2.3 Impact of Text-Only vs. SMILES-Only vs. Multi-Modal Querying

Table 6.4 presents retrieval performance across query modalities and pseudo-positive loss weights α, with temperature and per-batch missing ratio held fixed at $\tau = 0.20$ and 50% respectively. Three query modalities are evaluated: text-only, SMILES-only, and multi-modal, allowing direct assessment of the information contributed by each modality to the retrieval objective.

Table 6.4: Impact of query modality on retrieval performance across pseudo-positive loss weight α values, at fixed $\tau = 0.20$ and 50% per-batch missing ratio. Results are reported for text-only, SMILES-only, and multi-modal queries. The control uses random in-batch SMILES as the pseudo-positive with $\alpha = 1.0$.

		Text query		**SMILES query**		**Multi-modal query**	
Condition	$\boldsymbol{\alpha}$	**R'@10**	**nDCG'@10**	**R'@10**	**nDCG'@10**	**R'@10**	**nDCG'@10**
Random negative (control)	1.0	0.167	0.211	0.479	0.434	0.520	0.460
Hard negative	0.1	**0.172**	**0.231**	0.436	0.395	0.503	0.434
	0.5	0.157	0.207	0.470	0.422	0.528	0.451
	0.8	0.149	0.199	0.494	0.443	0.530	0.460
	1.0	0.161	0.204	**0.509**	**0.449**	**0.547**	**0.467**
	2.0	0.169	0.199	0.495	0.441	0.534	0.461

Effect of query modality. The most striking finding in Table 6.4 is the large performance gap between text-only and SMILES-only queries across all α values. At $\alpha = 1.0$, the hard negative condition achieves R'@10= 0.161 and nDCG'@10= 0.204 for text queries, compared to R'@10= 0.509 and nDCG'@10= 0.449 for SMILES queries — a difference of over 0.34 on R'@10. This result has a meaningful chemical interpretation. Traditional chemical information retrieval systems such as Reaxys and SciFinder have historically been queried predominantly through chemical names, synonyms, and textual descriptions, reflecting the way chemists naturally communicate about molecules. Although they do support structure search through SMILES, their fusion of results is mostly based on boolean logic to rank overlaps higher than others without any joint context between the two modes of retrieval. Furthermore, they are strictly keyword-based and do not have semantic knowledge of chemical, reactions or their properties. Additionally, natural language descriptions of chemical entities are inherently ambiguous: the same molecule may be referred to by its IUPAC name, common name, trade name, or a partial structural description, depending on the source document and author convention. SMILES strings, by contrast, provide an unambiguous, structure-first representation that explicitly encodes atomic connectivity, stereochemistry, and bonding. The model, having been trained contrastively on passage–SMILES pairs, learns to align SMILES embeddings with their chemical semantics in a way that captures implicit structural constraints — aromaticity, functional group identity, substructure relationships — that are only loosely and inconsistently conveyed in natural language. The superiority of SMILES queries therefore reflects not a deficiency of the text encoder, but the fundamental information-theoretic advantage of structure-based query representations for chemical retrieval tasks. This finding suggests that equipping chemists with structure-based query interfaces, even in text-centric document

retrieval settings, would substantially improve their ability to locate relevant passages in patent and publication archives.

Multi-modal queries and the role of α. The multi-modal query condition consistently achieves the strongest retrieval performance across all α values, with the best result of R'@10= 0.547 and nDCG'@10= 0.467 at $\alpha = 1.0$. This outperforms both text-only and SMILES-only queries at the same α, confirming that the two modalities contribute complementary information to the retrieval objective. The SMILES query provides precise structural grounding, while the text query supplies contextual chemical knowledge — reaction conditions, biological activity, synthetic utility — that is not encoded in the molecular graph itself. Their fusion via Reciprocal Rank Fusion therefore combines the structural precision of SMILES-based retrieval with the contextual breadth of text-based retrieval, producing a multi-modal query interface that more closely mirrors the way a chemist reasons about a molecule of interest.

Across all three query modalities, the random negative control at $\alpha = 1.0$ is outperformed by the hard negative condition at $\alpha = 1.0$ on SMILES and multi-modal queries, confirming that hard negative pseudo-positive supervision produces superior embedding alignment for structure-informed retrieval. The one exception is the text-only query, where the hard negative condition achieves its best performance not at $\alpha = 1.0$ but at the lower weight $\alpha = 0.1$, recording R'@10= 0.172 and nDCG'@10= 0.231. This suggests that for text-only retrieval, a small pseudo-positive contribution is sufficient and that higher α values introduce gradient signal that disproportionately shapes the embedding space toward structural alignment at the expense of purely textual semantic relationships. This is consistent with the broader finding that the optimal α is modality-dependent, and that the pseudo-positive objective primarily benefits structure-informed query modes where the SMILES encoder's representational precision is leveraged at retrieval time.

6.2.4 Statistical Testing (t-test)

Tables 6.5 and 6.6 present paired t-test results evaluating the statistical robustness of the observed retrieval improvements under two different baseline conditions. Together, these two tests allow us to disentangle the contribution of query modality from the contribution of the pseudo-positive training strategy.

Table 6.5: **Effect of Query Modality:** Paired t-test results comparing hard negative multi-modal query conditions against the random negative baseline (Control: $\alpha = 1.0$, random negatives, text-only retrieval) (Treatments: Hard-Negative, Text+SMILES retrieval)

Metric	α	t**-stat**	p**-raw**	p**-Bonferroni**	**Significant**
R'@10	0.1	-6.925	7.68×10^{-8}	6.14×10^{-7}	Yes
nDCG'@10	0.1	-4.390	1.16×10^{-4}	9.25×10^{-4}	Yes
R'@10	0.5	-7.299	2.69×10^{-8}	2.15×10^{-7}	Yes
nDCG'@10	0.5	-4.596	6.41×10^{-5}	5.13×10^{-4}	Yes
R'@10	0.8	-6.844	9.66×10^{-8}	7.72×10^{-7}	Yes
nDCG'@10	0.8	-4.608	6.19×10^{-5}	4.96×10^{-4}	Yes
R'@10	1.0	-7.347	2.36×10^{-8}	1.89×10^{-7}	Yes
nDCG'@10	1.0	-4.842	3.14×10^{-5}	2.51×10^{-4}	Yes

Table 6.6: **Effect of Pseudo-Positive Pairings:** Control and treatments both have the same query modalities but uses the random negative pairing (Control: $\alpha = 1.0$, random negatives, Text+SMILES retrieval) (Treatments: Hard-Negative, Text+SMILES retrieval)

Metric	α	t**-stat**	p**-raw**	p**-Bonferroni**	**Significant**
R'@10	0.1	0.485	0.631	1.000	No
nDCG'@10	0.1	0.785	0.438	1.000	No
R'@10	0.5	-0.491	0.627	1.000	No
nDCG'@10	0.5	0.982	0.333	1.000	No
R'@10	0.8	-0.762	0.451	1.000	No
nDCG'@10	0.8	0.055	0.956	1.000	No
R'@10	1.0	-1.952	0.060	0.478	No
nDCG'@10	1.0	-0.964	0.342	1.000	No

Query modality is the dominant driver of improvement: When the random negative text-only baseline is used as the control (Table 6.5), all hard negative multi-modal conditions produce statistically significant improvements on both R'@10 and nDCG'@10 across all α values, with Bonferroni-corrected p-values well below 0.001. The t-statistics are large in magnitude (ranging from -4.39 to -7.35), indicating that the improvement is not only statistically reliable but highly consistent across the 33 queries (only queries with at least one candidate scored above 0) in the test

collection. This result confirms that the combination of hard negative pseudo-positive supervision with multi-modal Text+SMILES querying produces a robust and reproducible retrieval improvement over a text-only random negative baseline. The primary driver of this improvement, however, is the query modality rather than the training strategy: as shown in Table 6.4, switching from text-only to Text+SMILES querying alone produces a large gain in R'@10 (from approximately 0.167 to 0.520) even under the random negative condition, far exceeding any gain attributable to the choice of pseudo-positive selection strategy.

Effect of pseudo-positive training strategy is numerically present but not statistically significant: When the random negative multi-modal baseline is used as the control (Table 6.6), holding query modality constant across both control and treatment, none of the hard negative conditions produce statistically significant differences after Bonferroni correction. This is an important and honest qualification of the results: while hard negative pseudo-positive supervision consistently produces numerically higher retrieval scores than random negative supervision at higher α values, as seen in Table 6.4, the magnitude of this difference does not reach statistical significance given the test collection size of $n = 35$ queries. The Bonferroni-corrected p-values are 1.0 for all but one comparison, reflecting both the conservatism of the correction and the limited statistical power available with a small expert-annotated test collection of this size. It is worth noting that at $\alpha = 1.0$, the R'@10 comparison between hard negative and random negative multi-modal conditions produces the lowest raw p-value in Table 6.6 ($p = 0.060$, Bonferroni $p = 0.478$), approaching but not crossing the conventional significance threshold. This is suggestive of a genuine directional effect of the pseudo-positive loss weight at its maximum value, and is consistent with the monotonically increasing performance trend observed across α values in Table 6.4. A larger test collection would likely provide sufficient statistical power to resolve this effect.

Implications and limitations: These results carry two practical implications. First, the most consequential design decision for chemical information retrieval in this framework is the availability of a SMILES query modality: providing chemists with structure-based query interfaces alongside textual queries produces a large, statistically robust improvement in retrieval quality that is independent of the training strategy. Second, hard negative pseudo-positive supervision provides a consistent numerical advantage over random pseudo-positive supervision, particularly at higher α values and for SMILES and multi-modal query modes, but this advantage cannot be confirmed as statistically significant with the current test collection. This does not mean the effect is absent — the numerical trends are consistent and directionally clear across all conditions. Constructing a

Table 6.7: Summary of Best Performing Models From Our Experiments

Condition	Missing %	τ	α	R'@1	R'@5	R'@10	nDCG'@1	nDCG'@5	nDCG'@10
Standard InfoNCE	0% (no missing)	0.20	0.1	0.093	**0.354**	0.477	**0.379**	**0.422**	0.451
Random Negative	50%	0.10	1.0	0.093	0.320	0.520	0.288	0.375	0.460
Hard Negative	10%	0.10	1.0	**0.100**	0.339	**0.533**	0.308	0.398	**0.470**

larger graded relevance collection with more queries would be a natural next step to definitively establish the statistical significance of the pseudo-positive training contribution independently of the query modality effect.

6.3 Summary

This chapter investigated the impact of pseudo-positive contrastive supervision on cross-modal chemical information retrieval, evaluating the interplay between per-batch missing sample ratio, temperature τ, pseudo-positive loss weight α, pseudo-positive selection strategy, and query modality across a graded relevance test collection of 35 expert-annotated queries. Table 6.7 summarises the best performing configuration from each experimental condition. The results provide evidence across several dimensions regarding the nature of pseudo-positive supervision and its interaction with training data volume, temperature, and query modality.

A notable result concerns the hard negative condition at 50% per-batch missing ratio ($\tau = 0.10$, $\alpha = 1.0$), which achieves R'@1= 0.100, R'@10= 0.533, and nDCG'@10= 0.470, outperforming the standard InfoNCE control on R'@1, nDCG'@1, nDCG'@5, and nDCG'@10 despite a substantial difference in training data volume. The standard InfoNCE control trains on the full 336K high-fidelity passage–SMILES pairs, while the hard negative condition at 50% missing sees only 172K valid pairs per epoch — approximately half the paired supervision budget — supplemented by 44K ungrounded passages assigned hard negative pseudo-positives. That this condition matches or exceeds the data-rich control on ranking quality metrics (nDCG) while remaining competitive on recall (R'@10: 0.533 vs. 0.553) suggests that hard negative pseudo-positive supervision contributes representational enrichment that is not straightforwardly replicated by additional high-fidelity pairs alone. This is an encouraging finding: ungrounded contextual passages, when coupled with hard negative pseudo-positive alignment, appear to provide a complementary supervision signal that improves ranking precision beyond what the standard InfoNCE objective achieves with twice the paired data, though we note that the test collection size of $n = 35$ limits the strength of conclusions

that can be drawn from individual metric comparisons.

A further observation that supports the role of regularization in this framework concerns the best performing standard InfoNCE configuration, which achieves its highest nDCG'@1, nDCG'@5 and R'@5 scores at $\tau = 0.20$ rather than $\tau = 0.10$. This is notable because a higher temperature flattens the softmax distribution over negatives, assigning more uniform weights across the negative set and thereby reducing the penalty attributed to the hardest negatives relative to easier ones. In effect, $\tau = 0.20$ acts as an implicit regularizer on the standard InfoNCE objective, preventing the model from over-concentrating its gradient signal on the few most confusable SMILES in the batch and instead distributing learning signal more evenly across the negative set. The fact that this softer temperature produces the best nDCG performance for the standard control, despite being suboptimal for the pseudo-positive conditions, is consistent with the broader finding that some form of regularization — whether through temperature softening, random negative pseudo-positives, or hard negative pseudo-positives at moderate missing ratios — is beneficial for the learned chemical embedding space. It further suggests that the improvements observed under pseudo-positive supervision are not solely attributable to the additional data coverage provided by ungrounded passages, but partly to the regularization effect that the auxiliary loss term introduces on top of the standard contrastive objective.

The random negative condition at 50% missing and $\tau = 0.10$ also produces competitive results (R'@10= 0.520, nDCG'@10= 0.460), which is consistent with the pseudo-positive loss acting as a regularizer that reduces overfitting to frequently occurring valid pair samples, rather than providing a meaningful chemical alignment signal in its own right. Hard negative pseudo-positive supervision produces numerically higher scores than random negative selection at $\alpha = 1.0$ across most metrics, though paired t-tests with Bonferroni correction confirm that this difference does not reach statistical significance at the current test collection size, and a larger evaluation collection would be needed to establish this effect with appropriate confidence.

The most statistically robust finding of the chapter is the contribution of query modality to retrieval performance: all hard negative multi-modal Text+SMILES query conditions are significantly superior to the random negative text-only baseline across all α values ($p < 0.001$ after Bonferroni correction), with the transition from text-only to multi-modal querying producing a gain of over 0.38 on R'@10. This gain substantially exceeds any effect attributable to the training strategy alone and has a natural chemical interpretation: SMILES representations provide an unambiguous, structure-first query modality that the model learns to align with chemical semantics during contrastive training, while text queries contribute contextual knowledge that complements structural

precision in the fused multi-modal retrieval setting. This combination more closely reflects how practising chemists reason about molecules of interest than the text-only querying paradigm of existing tools such as Reaxys and SciFinder. Taken together, the results suggest that pseudo-positive supervision from ungrounded chemical passages, combined with structure-informed multi-modal querying, offers a promising and data-efficient direction for cross-modal chemical information retrieval.

Chapter 7

Conclusion

This chapter summarizes the major contributions of this thesis and situates them within the broader landscape of chemical information retrieval and multi-modal representation learning. We revisit each research question in turn, assess the degree to which the experimental evidence supports the original hypothesis, and identify the limitations and most promising future directions arising from each contribution.

7.1 Chemical Extraction, Linking and Indexing System

A foundational contribution of this work is the development of a complete end-to-end indexing pipeline for chemical patent literature, capable of extracting text passages at page-level granularity, identifying chemical entity mentions, converting them to canonical SMILES representations via OPSIN and external knowledge bases, and linking extracted passages to their associated molecular diagrams anywhere within the source PDF. The result is a richly structured index in which every passage and diagram carries provenance metadata — source document, page number, and spatial location — enabling chemists to navigate seamlessly between textual descriptions and molecular representations within long patent documents. Critically, improvements to the extraction pipeline yielded a greater than 100-fold increase in the number of passages with positively paired SMILES samples relative to our preliminary work, providing sufficient data to train large-scale cross-modal retrieval models. The final corpus comprises 373K valid passage–SMILES pairs and 44K ungrounded passages, both of which are leveraged in the contrastive training framework developed in subsequent chapters.

Limitations and future work. The extraction pipeline, while substantially improved, remains fragmented: separate subsystems handle text segmentation, chemical named entity recognition, SMILES conversion, and diagram parsing, and errors introduced at any stage cascade through the remainder of the pipeline, reducing the total number of recoverable valid pairs. A unified passage detection, extraction, and chemical entity identification system would not only reduce this error propagation but improve interpretability and allow end-to-end optimisation of the full extraction process. The molecular diagram parser additionally struggles with complex diagram types, particularly those containing R-groups representing wildcard substituents and hyperplane representations of molecular geometry. Extending support to these diagram classes would meaningfully increase the coverage of the indexed collection and reduce reliance on manual intervention. Improving the extraction foundation before the retrieval layer is trained would also directly benefit retrieval quality, since more complete and accurate passage–molecule links reduce the probability of missing chemically relevant context in the first place.

7.2 Pooling and Test Collection

A second major contribution of this work is the construction of the first graded relevance test collection for chemical information retrieval over scientific document data, developed in collaboration with expert chemistry faculty and senior doctoral researchers from multiple universities. The collection comprises 35 multi-modal queries with 10–30 expert-annotated candidates per query, pooled from a diverse ensemble of text and molecular retrieval models and fused using a bucketing-based Reciprocal Rank Fusion strategy that prioritises candidates with evidence from both modalities. A novel relevance framework was developed specifically for the document-page retrieval context, recognising that information relevant to a given query is often distributed across neighbouring pages rather than concentrated in a single passage. Under this framework, a candidate is considered completely non-relevant only if neither the candidate nor its immediate page neighbourhood (± 1 page) contains any information pertinent to the query's information need. This operationalisation more faithfully reflects the navigation task chemists perform when searching long patent documents and avoids penalising retrieval systems for surfacing candidates that are contextually adjacent to the ground truth.

Limitations and future work. The pooling process, while designed to be as diverse as possible, relied predominantly on keyword-based and early-stage neural models that lack implicit chemical property knowledge. The candidate pool would benefit from the inclusion of more diverse dense re-

trieval models, including transformer-based and graph-based molecular encoders trained on different chemical corpora, which would increase the likelihood of surfacing chemically relevant candidates that keyword-based systems miss. The most consequential limitation, however, is the test collection size: with only 35 annotated queries, the statistical power available for distinguishing training strategy effects from query modality effects is insufficient, as demonstrated by the significance tests reported in Chapter **??**. Expanding the collection to 100 or more queries, spanning a broader range of chemical query types including reaction-centric, scaffold-based, and mechanism-based queries, would substantially increase statistical power and enable more fine-grained analysis of retrieval performance across chemically distinct information needs.

7.3 Hard-Negative Based Pseudo-Positive Objective Function

The central methodological contribution of this thesis is a hybrid contrastive training objective that jointly leverages valid passage–SMILES pairs and ungrounded chemical passages within a single InfoNCE-based training framework. The key insight motivating this approach is that ungrounded passages — those containing chemical language but no extractable molecular entity mentions — are not uninformative noise to be discarded, but contextually rich descriptions of chemical processes, reaction conditions, and synthetic rationale that can enrich the learned embedding space without requiring ground-truth SMILES supervision. By assigning the hardest in-batch negative SMILES as a pseudo-positive for each ungrounded passage and weighting its contribution through a scalar α, the proposed objective introduces a regularisation effect that prevents the embedding space from collapsing around the most frequently occurring valid pair samples while providing additional textual coverage for general chemical concepts that apply across a broad range of molecular structures.

Empirical results demonstrate that a per-batch missing ratio of 10% with $\alpha = 1.0$ and $\tau = 0.10$ yields the strongest retrieval performance, achieving R'@10= 0.547 and nDCG'@10= 0.467 under multi-modal Text+SMILES querying — outperforming the random negative baseline on both metrics. Statistical significance testing further establishes that the dominant driver of retrieval improvement is the query modality rather than the training strategy: all hard negative multi-modal conditions are significantly superior to the random negative text-only baseline ($p < 0.001$ after Bonferroni correction), while the difference between hard negative and random negative training at matched query modality does not reach statistical significance at the current test collection size, though directionally consistent numerical trends at $\alpha = 1.0$ suggest a genuine effect that a larger evaluation collection would likely confirm. These results demonstrate that equipping chemists with

structure-based SMILES query interfaces, rather than text-only interfaces analogous to existing tools such as Reaxys and SciFinder, is the most consequential design decision for practical chemical retrieval systems, and that the pseudo-positive training objective provides a complementary and consistent, if modest, additional benefit.

Limitations and future work. Several directions present themselves for extending the pseudo-positive training framework. The most immediate improvement would be the introduction of a momentum queue that maintains SMILES embeddings from recent batches, allowing hard negative mining to draw from a much larger and more chemically diverse candidate pool than the current in-batch selection, which is limited to 30 candidates. A curriculum learning schedule that begins with zero ungrounded passages and gradually increases the missing ratio as the encoder develops chemical alignment from valid pairs would address the early-training instability observed in the positive similarity curves, ensuring that pseudo-positive assignments are always made against a sufficiently calibrated encoder. Additionally, soft pseudo-positive labelling via Tanimoto similarity weighting — assigning a continuous target proportional to the structural similarity between the hard negative and other batch SMILES rather than treating it as a binary positive — would reduce the gradient conflict that arises when the pseudo-positive is only weakly chemically related to the ungrounded passage. Finally, a self-training approach using the clean model to generate and filter pseudo-SMILES labels for ungrounded passages prior to joint training would convert noisy geometric pseudo-positives into curated soft labels grounded in the model's own learned chemical representations.

7.4 Concluding Thoughts

This work explores several complementary dimensions of chemical information retrieval, collectively advancing our understanding of how chemical information needs can be represented, extracted, and effectively satisfied at scale. Beyond improving retrieval performance in a narrow sense, the proposed methods highlight the structure and complexity of chemical knowledge embedded in scientific documents, and the challenges involved in faithfully capturing it.

A key outcome of this study is the set of contributions that extend beyond the immediate application domain of chemistry. In particular, the modeling strategies developed here suggest broader applicability to other large-scale, weakly supervised, and noise-prone data environments, such as web corpora, social media streams, and large scientific archives. In these settings, obtaining fully cu-

rated or noise-free supervision is often infeasible or prohibitively expensive, making semi-supervised and weakly supervised learning paradigms especially relevant.

This work also motivates a reconsideration of traditional notions of relevance in information retrieval. In domains where information is fragmented, distributed across multiple modalities, or only partially observable, conventional binary or document-level relevance assumptions may be insufficient. Our findings highlight the need for more nuanced and graded formulations of relevance that better reflect real-world information structure in a document page-level based retrieval setting.

Finally, we believe this work establishes a foundation for future research in large-scale semi-supervised retrieval systems. The ideas introduced here—particularly around pseudo-positive supervision and multi-modal alignment—provide reusable building blocks that can be extended, refined, and adapted across scientific and industrial search settings. We hope this contributes to a broader shift toward more flexible, data-efficient retrieval frameworks capable of operating under realistic annotation constraints.

Chapter 8

List of Publications

Table 8.1: List of Papers published before proposal defense in reverse chronological order

No.	Title	Venue	Year
1.	*Targeted Multi-Modal Passage Search for Molecules and their Synthesis Pathways (Demo)*, **Abhisek Dey**, Nathaniel Stanley, Richard Zanibbi	SIGIR	2025
2.	*Multimodal Search in Chemical Documents and Reactions (Demo)*, Ayush Kumar Shah[1], **Abhisek Dey**[1], Leo Luo, Bryan Amador, Patrick Phillipy, Ming Zhong, Siru Ouyang, David Friday, David Bianchi, Nick Jackson, Richard Zanibbi, Jiawei Han	SIGIR	2025
3.	*MoleculeMiner: Extracting and Linking Molecule Figures with Tabular Metadata (Demo)*, **Abhisek Dey**, Nathaniel Stanley	IJCAI	2025
4.	*ChemScraper: Leveraging PDF graphics instructions for molecular diagram parsing*, Ayush Kumar Shah, Bryan Amador, **Abhisek Dey**, Ming Creekmore, Blake Ocampo, Scott Denmark, Richard Zanibbi	IJDAR	2024
5.	*Searching the ACL Anthology with Math Formulas and Text (Demo)*, Bryan Amador, Matt Langsenkamp, **Abhisek Dey**, Ayush Kumar Shah, Richard Zanibbi	SIGIR	2023
6.	*ScanSSD-XYc: Faster Detection for Math Formulas*, **Abhisek Dey** and Richard Zanibbi	GREC	2021
7.	*A Math Formula Extraction and Evaluation Framework for PDF Documents*, Ayush Kumar Shah, **Abhisek Dey**, Richard Zanibbi	ICDAR	2021

Bibliography

[1] Anish Acharya, Sujay Sanghavi, Li Jing, Bhargav Bhushanam, Dhruv Choudhary, Michael Rabbat, and Inderjit Dhillon. Positive unlabeled contrastive learning. *arXiv preprint arXiv:2206.01206*, 2022.

[2] Tagir Akhmetshin, Arkadii I. Lin, Daniyar Mazitov, Evgenii Ziaikin, Timur Madzhidov, and Alexandre Varnek. ZINC 250K data sets. 12 2021.

[3] Saber A. Akhondi, Hinnerk Rey, Markus Schwörer, Michael Maier, John Toomey, Heike Nau, Gabriele Ilchmann, Mark Sheehan, Matthias Irmer, Claudia Bobach, Marius Doornenbal, Michelle Gregory, and Jan A. Kors. Automatic identification of relevant chemical compounds from patents. *Database*, 2019:1–14, 2019.

[4] Bryan Amador, Matt Langsenkamp, Abhisek Dey, Ayush Kumar Shah, and Richard Zanibbi. Searching the acl anthology with math formulas and text. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 3110–3114, New York, NY, USA, 2023. Association for Computing Machinery.

[5] Hangbo Bao, Li Dong, and Furu Wei. M3ae: Multimodal masked autoencoders for unified modality learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[6] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems*, 35, 2022.

[7] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th*

International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3615–3620, Hong Kong, China, November 2019. Association for Computational Linguistics.

[8] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv:2004.05150*, 2020.

[9] Syed Saqib Bukhari, Zaryab Iftikhar, and Andreas Dengel. Chemical structure recognition (csr) system: Automatic analysis of 2d chemical structures in document images. *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, pages 1432–1437, 2019.

[10] Yinqiong Cai, Yueyuan Cao, Yixing Li, Liang Liu, Jiafeng Guo, and Xueqi Cheng. Hard negatives or false negatives: Correcting pooling bias in training neural ranking models. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management (CIKM 2022)*, pages 118–127. ACM, 2022.

[11] Daniel Campos and Heng Ji. Img2smi: Translating molecular structure images to simplified molecular-input line-entry system. pages 1–12, 2021.

[12] David Campos, Sérgio Matos, and José L. Oliveira. A document processing pipeline for annotating chemical entities in scientific documents. *Journal of Cheminformatics*, 7:1–10, 2015.

[13] Adrià Cereto-Massagué, María José Ojeda, Cristina Valls, Miquel Mulero, Santiago Garcia-Vallvé, and Gerard Pujadas. Molecular fingerprint similarity search in virtual screening. *Methods*, 71:58–63, 2015.

[14] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E Hinton. A simple framework for contrastive learning of visual representations. bt - proceedings of the 37th international conference on machine learning, icml 2020, 13-18 july 2020, virtual event. *Icml*, pages 1597–1607, 2020.

[15] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in Neural Information Processing Systems*, 2020-Decem:1–18, 2020.

[16] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1–9, 2015.

[17] Yen Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12375 LNCS:104–120, 2020.

[18] Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta: Large-scale self-supervised pretraining for molecular property prediction. 2020.

[19] Ching-Yao Chuang, Joshua Robinson, Lin Yen-Chen Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. In *NeurIPS*, 2020.

[20] Paolo Comelli, Paolo Ferragina, Mario Notturno Granieri, and Flavio Stabile. Optical recognition. 44:627–631, 1995.

[21] Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, page 758–759, New York, NY, USA, 2009. Association for Computing Machinery.

[22] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. Overview of the TREC 2020 deep learning track. In *Proceedings of the 29th Text REtrieval Conference (TREC 2020)*. NIST, 2021.

[23] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. Overview of the TREC 2021 deep learning track. In *Proceedings of the 30th Text REtrieval Conference (TREC 2021)*. NIST, 2022.

[24] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M. Voorhees. Overview of the TREC 2019 deep learning track. In *Proceedings of the 28th Text REtrieval Conference (TREC 2019)*. NIST, 2020.

[25] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M. Voorhees. Overview of the TREC 2022 deep learning track. In *Proceedings of the 31st Text REtrieval Conference (TREC 2022)*. NIST, 2023.

[26] Jingyi Cui, Weiran Huang, Yifei Wang, and Yisen Wang. Rethinking weak supervision in helping contrastive learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 6448–6467. PMLR, 23–29 Jul 2023.

[27] Giulio Degtyarenko, Janna Hastings, Pedro de Matos, Michael Ennis, Marcus C Chibucos, Alan McNaught, Jie Zhang, Neil J Williams, Gareth Owen, Evan Bolton, Antony J Williams, et al. The chebi reference database and ontology for biologically relevant chemistry: enhancements for 2016. *Nucleic Acids Research*, 44(D1):D1214–D1219, 2016.

[28] Kirill Degtyarenko, Paula De matos, Marcus Ennis, Janna Hastings, Martin Zbinden, Alan Mcnaught, Rafael Alcántara, Michael Darsow, Mickaël Guedj, and Michael Ashburner. Chebi: A database and ontology for chemical entities of biological interest. *Nucleic Acids Research*, 36, 1 2008.

[29] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255. IEEE, 2009.

[30] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova Google, and A I Language. Bert: Pre-training of deep bidirectional transformers for language understanding. *Naacl-Hlt 2019*, pages 4171–4186, 2018.

[31] Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1:4171–4186, 2019.

[32] Abhisek Dey and Nathaniel Stanley. Moleculeminer: Extracting and linking molecule figures with tabular metadata. In *International Joint Conference on Artificial Intelligence (IJCAI)*, volume 34, 2025.

[33] Abhisek Dey, Nathaniel Stanley, and Richard Zanibbi. Targeted multi-modal passage search for molecules and their synthesis pathways. In *International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR) (To Appear)*, volume 48, 2025.

[34] Abhisek Dey and Richard Zanibbi. *ScanSSD-XYc: Faster Detection for Math Formulas*, volume 12916 LNCS. Springer International Publishing, 2021.

[35] Siying Dong, Andrew Kryczka, Yanqin Jin, and Michael Stumm. Rocksdb: Evolution of development priorities in a key-value store serving large-scale applications. *ACM Trans. Storage*, 17(4), October 2021.

[36] Carl Edwards, Cheng Xiang Zhai, and Heng Ji. Text2mol: Cross-modal molecule retrieval with natural language queries. *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, pages 595–607, 2021.

[37] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. In *British Machine Vision Conference (BMVC)*, 2018.

[38] Vincent Fan, Yujie Qian, Alex Wang, Amber Wang, Connor W. Coley, and Regina Barzilay. Openchemie: An information extraction toolkit for chemistry literature. *Journal of Chemical Information and Modeling*, 64:5521–5534, 7 2024.

[39] Igor V. Filippov and Marc C. Nicklaus. Optical structure recognition software to recover chemical information: Osra, an open source solution. *Journal of Chemical Information and Modeling*, 49:740–743, 2009.

[40] Andrea Frome, Greg S Corrado, Jonathon Shlens, Samy Bengio Jeffrey Dean, Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems (NeurIPS), 2013*, 2013.

[41] Stephen Walter Gabrielson. Scifinder, 2018.

[42] Rohit Girdhar, Alaaeldin El-Nouby, and Zhuang Liu Mannat Singh Kalyan Vasudev Alwala Armand Joulin Ishan Misra. Imagebind: One embedding space to bind them all. In *CVPR*, 2023.

[43] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing XU, and Yunhe Wang. Transformer in transformer. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, pages 15908–15919, 2021.

[44] David R. Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

[45] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[46] Stephen Heller. Inchi – the worldwide chemical structure standard. *Journal of Cheminformatics*, 6:1–9, 2014.

[47] Stephen R. Heller, Alan McNaught, Igor Pletnev, Stephen Stein, and Dmitrii Tchekhovskoi. *InChI, the IUPAC International Chemical Identifier*, volume 7. Journal of Cheminformatics, 2015.

[48] K. M. Hettne, R. H. Stierum, M. J. Schuemie, P. J. Hendriksen, B. J. Schijvenaars, E. M. van Mulligen, J. Kleinjans, and J. A. Kors. A dictionary to identify small molecules and drugs in free text. *Bioinformatics*, 25(22):2983–2991, November 2009. Epub 2009 Sep 16.

[49] P. Ibison, M. Jacquot, F. Kam, A. G. Neville, R. W. Simpson, C. Tonnelier, T. Venczel, and A. P. Johnson. Chemical literature data extraction: The clide project. *Journal of Chemical Information and Computer Sciences*, 33:338–344, 1993.

[50] Aashi Jain, Mandy Guo, Krishna Srinivasan, Ting Chen, Sneha Kudugunta, Chao Jia, Yinfei Yang, and Jason Baldridge. Mural: Multimodal, multitask retrieval across languages. In *EMNLP*, pages 3449–3463, 2021.

[51] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

[52] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolov8, 2023.

[53] Kaggle. Bristol-Myers Squibb - molecular translation competition, 2021. Kaggle.

[54] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:664–676, 4 2017.

[55] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online, November 2020. Association for Computational Linguistics.

[56] Anthony Kay. Tesseract: an open-source optical character recognition engine. *Linux J.*, 2007(159):2, July 2007.

[57] Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. *SIGIR 2020 - Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 39–48, 2020.

[58] Dae Hyun Kim, Enamul Hoque, Juho Kim, and Maneesh Agrawala. Facilitating document reading by linking text and tables. In *UIST 2018 - Proceedings of the 31st Annual ACM*

Symposium on User Interface Software and Technology, pages 423–434. Association for Computing Machinery, Inc, 10 2018.

[59] Sangyeup Kim, Nayeon Kim, Yinhua Piao, and Sun Kim. Grapht5: Unified molecular graph–language modeling via multi-modal cross-token attention. *arXiv preprint arXiv:2503.07655*, mar 2025.

[60] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, Leonid Zaslavsky, Jian Zhang, and Evan E Bolton. Pubchem 2023 update. *Nucleic Acids Research*, 51(D1):D1373–D1380, 10 2022.

[61] Sunghwan Kim, Paul A Thiessen, Evan E Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A Shoemaker, Jiyao Wang, Bo Yu, Jian Zhang, and Stephen H Bryant. Pubchem substance and compound databases. *Nucleic acids research*, 44:D1202–13, 1 2016.

[62] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. *Proceedings of Machine Learning Research*, 139:5583–5594, 2021.

[63] Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. In *NeurIPS Deep Learning and Representation Learning Workshop*, 11 2014.

[64] Roman Klinger, Corinna Kolářik, Juliane Fluck, Martin Hofmann-Apitius, and Christoph M. Friedrich. Detection of iupac and iupac-like chemical names. *Bioinformatics*, 24:268–276, 2008.

[65] Nicholas Kong, Marti A. Hearst, and Maneesh Agrawala. Extracting references between text and charts via crowdsourcing. In *Conference on Human Factors in Computing Systems - Proceedings*, pages 31–40. Association for Computing Machinery, 2014.

[66] Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong Lu, Robert Leaman, Yanan Lu, Donghong Ji, Daniel M. Lowe, Roger A. Sayle, Riza Theresa Batista-Navarro, Rafal Rak, Torsten Huber, Tim Rocktäschel, Sérgio Matos, David Campos, Buzhou Tang, Hua Xu, Tsendsuren Munkhdalai, Keun Ho Ryu, S. V. Ramanan, Senthil Nathan, Slavko Žitnik, Marko Bajec, Lutz Weber, Matthias Irmer, Saber A. Akhondi, Jan A. Kors, Shuo Xu, Xin An, Utpal Kumar Sikdar, Asif Ekbal, Masaharu Yoshioka, Thaer M. Dieb, Miji Choi, Karin Verspoor, Madian Khabsa, C. Lee Giles, Hongfang Liu, Komandur Elayavilli Ravikumar, Andre Lamurias, Francisco M. Couto, Hong Jie Dai, Richard Tzong Han Tsai,

Caglar Ata, Tolga Can, Anabel Usié, Rui Alves, Isabel Segura-Bedmar, Paloma Martínez, Julen Oyarzabal, and Alfonso Valencia. The chemdner corpus of chemicals and drugs and its annotation principles. *Journal of Cheminformatics*, 7:1–17, 2015.

[67] Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. Combining language and vision with a multimodal skip-gram model. In *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*, pages 153–163. Association for Computational Linguistics, 2015.

[68] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part IV*, pages 212–228. Springer-Verlag, 2018.

[69] Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee. Multimodal prompting with missing modalities for visual recognition. *Computer Vision and Pattern Recognition (CVPR)*, 3 2023.

[70] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022.

[71] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX*, page 121–137, Berlin, Heidelberg, 2020. Springer-Verlag.

[72] Xue Li, Jiong Yu, Shaochen Jiang, Hongchun Lu, and Ziyang Li. Msvit: Training multiscale vision transformers for image retrieval. *IEEE Transactions on Multimedia*, 26:2809–2823, 2024.

[73] Yanchi Li, Guanyu Chen, and Xiang Li. Automated Recognition of Chemical Molecule Images Based on an Improved TNT Model. *Applied Sciences*, 12(2):680, 2022.

[74] Yongxin Li, Ying Cheng, Yaning Pan, Wen He, Qing Wang, Rui Feng, and Xiaobo Zhang. Semantic-aware hard negative mining for medical vision-language contrastive pretraining. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, page 3133–3142, New York, NY, USA, 2025. Association for Computing Machinery.

[75] Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *NeurIPS*, 2024.

[76] Ronghao Lin and Haifeng Hu. Missmodal: Increasing robustness to missing modality in multimodal sentiment analysis. *TACL*, 2023.

[77] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014.

[78] Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei Xiao, and Anima Anandkumar. Multi-modal molecule structure-text model for text-based retrieval and editing. pages 1–31, 2022.

[79] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *ICLR 2020*, 2019.

[80] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE International Conference on Computer Vision*, pages 9992–10002, 2021.

[81] Daniel M. Lowe, Peter T. Corbett, Peter Murray-Rust, and Robert C. Glen. Chemical name to structure: Opsin, an open source solution. *Journal of Chemical Information and Modeling*, 51:739–753, 2011.

[82] Daniel M. Lowe, Peter T. Corbett, Peter Murray-Rust, and Robert C. Glen. Chemical name to structure: Opsin, an open source solution. *Journal of Chemical Information and Modeling*, 51(3):739–753, 2011. PMID: 21384929.

[83] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. *ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks.* Curran Associates Inc., Red Hook, NY, USA, 2019.

[84] Man Luo, Zhiyuan Fang, Tejas Gokhale, Yezhou Yang, and Chitta Baral. End-to-end knowledge retrieval with multi-modal queries. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 1:8573–8589, 2023.

[85] Mihai Lupu, Jimmy Huang, Jianhan Zhu, and John Tait. Trec-chem: large scale chemical information retrieval evaluation at trec. *SIGIR Forum*, 43(2):63–70, December 2009.

[86] Christian Lülf, Denis Mayr Lima Martins, Marcos Antonio Vaz Salles, Yongluan Zhou, and Fabian Gieseke. Clip-branches: Interactive fine-tuning for text-image retrieval. In *SIGIR 2024 - Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2719–2723. Association for Computing Machinery, Inc, 7 2024.

[87] Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2022-June:18156–18165, 2022.

[88] Xiao Long Ma, Caiming Xiong, Wenhao Liu, and Song-Chun Zhu. Snli-ve: A natural language inference dataset for the visual entailment task. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1906–1914, 2019.

[89] Craig Macdonald and Nicola Tonellotto. Declarative experimentation ininformation retrieval using pyterrier. In *Proceedings of ICTIR 2020*, 2020.

[90] Gerald Maggiora, Martin Vogt, Dagmar Stumpfe, and Jürgen Bajorath. Molecular similarity in medicinal chemistry, 4 2014.

[91] Behrooz Mansouri and Zahra Jahedibashiz. Clarifying questions in math information retrieval. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '23, page 149–158, New York, NY, USA, 2023. Association for Computing Machinery.

[92] Juraj Mavračić, Callum J. Court, Taketomo Isazawa, Stephen R. Elliott, and Jacqueline M. Cole. Chemdataextractor 2.0: Autopopulated ontologies for materials science. *Journal of Chemical Information and Modeling*, 61:4280–4289, 9 2021.

[93] Joe R. McDaniel and Jason R. Balmuth. Kekule: Ocr-optical chemical (structure) recognition. *Journal of Chemical Information and Computer Sciences*, 32:373–378, 1992.

[94] David Mendez, Anna Gaulton, A Patrícia Bento, Jon Chambers, Marleen De Veij, Eloy Félix, María Paula Magariños, Juan F Mosquera, Prudence Mutowo, Michal Nowotka, María Gordillo-Marañón, Fiona Hunter, Laura Junco, Grace Mugumbate, Milagros Rodriguez-Lopez, Francis Atkinson, Nicolas Bosc, Chris J Radoux, Aldo Segura-Cabrera, Anne Hersey, and Andrew R Leach. Chembl: towards direct deposition of bioassay data. *Nucleic acids research*, 47(D1):D930—D940, January 2019.

[95] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, 2013. Presented at ICLR 2013.

[96] Zijun Min, Bingshuai Liu, Liang Zhang, Jia Song, Jinsong Su, Song He, and Xiaochen Bo. Exploring optimal transport-based multi-grained alignments for text-molecule retrieval. In *Proceedings - 2024 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2024*, pages 2317–2324. Institute of Electrical and Electronics Engineers Inc., 2024.

[97] Prasenjit Mitra, C. Lee Giles, Bingjun Sun, and Ying Liu. Chemxseer: a digital library and data repository for chemical kinetics. In *Proceedings of the ACM First Workshop on CyberInfrastructure: Information Management in EScience*, CIMS '07, page 7–10, New York, NY, USA, 2007. Association for Computing Machinery.

[98] H. L. Morgan. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *Journal of Chemical Documentation*, 5(2):107–113, May 1965.

[99] Lucas Morin, Martin Danelljan, Maria Isabel Agea, Ahmed Nassar, Valery Weber, Ingmar Meijer, Peter Staar, and Fisher Yu. Molgrapher: Graph-based visual recognition of chemical structures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19552–19561, October 2023.

[100] Lucas Morin, Martin Danelljan, Maria Isabel Agea, Ahmed Nassar, Valery Weber, Ingmar Meijer, Peter Staar, and Fisher Yu. MolGrapher: Graph-based visual recognition of chemical structures. *arXiv*, 2023.

[101] Lucas Morin, Martin Danelljan, Maria Isabel Agea, Ahmed Nassar, Valery Weber, Ingmar Meijer, Peter Staar, and Fisher Yu. Molgrapher: Graph-based visual recognition of chemical structures, 2023.

[102] Lucas Morin, Valéry Weber, Gerhard Ingmar Meijer, Fisher Yu, and Peter W.J. Staar. Patcid: an open-access dataset of chemical structures in patent documents. *Nature Communications*, 15, 12 2024.

[103] National Library of Medicine. Pmc open access subset. `https://pmc.ncbi.nlm.nih.gov/tools/openftlist/`, 2003. Bethesda (MD): National Library of Medicine. [cited 2025, June 10].

[104] Niki Nezakati, Md Kaykobad Reza, Ameya Patil, Mashhour Solh, and M. Salman Asif. Mmp: Towards robust multi-modal learning with masked modality projection. *arXiv*, 10 2024.

[105] Noel O'Boyle and Andrew Dalke. Deepsmiles: An adaptation of smiles for use in machine-learning of chemical structures. *ChemRxiv*, pages 1–9, 2018.

[106] Noel M. O'Boyle, Michael Banck, Craig A. James, Chris Morley, Tim Vandermeersch, and Geoffrey R. Hutchison. Open Babel: An open chemical toolbox. *Journal of Cheminformatics*, 3(1):33, 2011.

[107] 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. Surechembl: A large-scale, chemically annotated patent document database. *Nucleic Acids Research*, 44:D1220–D1228, 2016.

[108] Jun-Hyung Park, Yeachan Kim, Mingyu Lee, Hyuntae Park, and SangKeun Lee. Moltres: Improving chemical language representation learning for molecular property prediction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14241–14254, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[109] Vikas Paruchuri and Datalab Team. Surya: A lightweight document ocr and analysis toolkit. `https://github.com/VikParuchuri/surya`, 2025. GitHub repository.

[110] Harry E. Pence and Antony Williams. Chemspider: An online chemical information resource. *Journal of Chemical Education*, 87(11):1123–1124, 2010.

[111] Florina Piroi, Mihai Lupu, Allan Hanbury, Walid Magdy, Alan P. Sexton, and Igor Filippov. Clef-ip 2012: Retrieval experiments in the intellectual property domain. *CEUR Workshop Proceedings*, 1178, 2012.

[112] Bryan A. Plummer, Liwei Wang, Chris Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2641–2649, 2015.

[113] Daniil Polykovskiy, Alexander Zhebrak, Benjamin Sanchez-Lengeling, Sergey Golovanov, Oktai Tatanov, Stanislav Belyaev, Rauf Kurbanov, Aleksey Artamonov, Vladimir Aladinskiy, Mark Veselov, Artur Kadurin, Simon Johansson, Hongming Chen, Sergey Nikolenko, Alan Aspuru-Guzik, and Alex Zhavoronkov. Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. *Frontiers in Pharmacology*, 2020.

[114] Yujie Qian, Jiang Guo, Zhengkai Tu, Zhening Li, Connor W. Coley, and Regina Barzilay. Molscribe: Robust molecular structure recognition with image-to-graph generation. *Journal of Chemical Information and Modeling*, 63:1925–1934, 2023.

[115] Yujie Qian, Jiang Guo, Zhengkai Tu, Zhening Li, Connor W. Coley, and Regina Barzilay. MolScribe: Robust molecular structure recognition with image-to-graph generation. *Journal of Chemical Information and Modeling*, 63(7):1925–1934, 2023.

[116] Yansheng Qiu, Ziyuan Zhao, Hongdou Yao, Delin Chen, and Zheng Wang. Modal-aware visual prompting for incomplete multi-modal brain tumor segmentation. In *MM 2023 - Proceedings of the 31st ACM International Conference on Multimedia*, pages 3228–3239. Association for Computing Machinery, Inc, 10 2023.

[117] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *Proceedings of Machine Learning Research*, 139:8748–8763, 2021.

[118] Kohulan Rajan, Achim Zielesny, and Christoph Steinbeck. Decimer: towards deep learning for chemical image recognition. *Journal of Cheminformatics*, 12:1–9, 2020.

[119] Kohulan Rajan, Achim Zielesny, and Christoph Steinbeck. Stout: Smiles to iupac names using neural machine translation. *Journal of Cheminformatics*, 13, 12 2021.

[120] Nikhil Rasiwasia, Jose Costa Periera, and Emanuale Coviello. A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM international conference on Multimedia*. ACM Digital Library, 2013.

[121] Louis C. Ray and Russell A. Kirsch. Finding chemical records by digital computers. *Science*, 126:814–819, 1957.

[122] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 91–99, 2015.

[123] Kirk Roberts, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, Kyle Lo, Ian Soboroff, Ellen Voorhees, Lucy Lu Wang, and William R. Hersh. TREC-COVID: Rationale and structure of an information retrieval shared task for COVID-19. *Journal of the American Medical Informatics Association*, 27(9):1431–1436, 2020.

[124] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389, April 2009.

[125] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. Okapi at trec-3. In *Text Retrieval Conference*, 1994.

[126] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241, 2015.

[127] Noureddin M. Sadawi, Alan P. Sexton, and Volker Sorge. Molrec at clef 2012 — overview and analysis of results. *CEUR Workshop Proceedings*, 1178, 2012.

[128] Nikunj Saunshi, Jordan T Ash, Surbhi Goel, Dipendra Misra, Cyril Zhang, Sanjeev Arora, Sham Kakade, and Akshay Krishnamurthy. Understanding contrastive learning requires incorporating inductive biases. In *39th International Conference on Machine Learning (ICML)*, 2022.

[129] Ayush Kumar Shah, Bryan Amador, Abhisek Dey, Ming Creekmore, Blake Ocampo, Scott Denmark, and Richard Zanibbi. Chemscraper: leveraging pdf graphics instructions for molecular diagram parsing. *International Journal on Document Analysis and Recognition (IJDAR)*, 27(3):395–414, Sep 2024.

[130] Zejiang Shen, Ruochen Zhang, Melissa Dell, Benjamin Charles Germain Lee, Jacob Carlson, and Weining Li. Layoutparser: A unified toolkit for deep learning based document image analysis. *arXiv preprint arXiv:2103.15348*, 2021.

[131] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015.

[132] Stanislaw Skonieczny. The iupac rules for naming organic molecules. *Journal of Chemical Education*, 83:1633–1637, 2006.

[133] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

[134] Xinying Song, Alex Salcianu, Yang Song, Dave Dopson, and Denny Zhou. Fast wordpiece tokenization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2089–2103. Association for Computational Linguistics, November 2021.

[135] Christopher Southan and Andras Stracz. Extracting and connecting chemical structures from text sources using chemicalize.org. *Journal of Cheminformatics*, 5:1–11, 2013.

[136] Joshua Staker, Kyle Marshall, Robert Abel, and Carolyn M. McQuaw. Molecular structure extraction from documents using deep learning. *Journal of Chemical Information and Modeling*, 59:1017–1029, 2019.

[137] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, Florence, Italy, July 2019. Association for Computational Linguistics.

[138] Matthew C. Swain and Jacqueline M. Cole. Chemdataextractor: A toolkit for automated extraction of chemical information from the scientific literature. *Journal of Chemical Information and Modeling*, 56:1894–1904, 10 2016.

[139] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019.

[140] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogenous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

[141] Anabel Usié, Rui Alves, Francesc Solsona, Miguel Vázquez, and Alfonso Valencia. Chener: chemical named entity recognizer. *Bioinformatics*, 30(7):1039–1040, 11 2013.

[142] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[143] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5998–6008, 2017.

[144] Jianxin Wang, Junyao Xiang, Yichao Quan, Qianqian Lu, and Jianmin Ma. Molecular property prediction by contrastive learning with attention-guided positive sample selection. *Bioinformatics*, 39(5):btad258, 2023.

[145] Yan Wang, Ruochi Zhang, Shengde Zhang, Liming Guo, Qiong Zhou, Bowen Zhao, Xiaotong Mo, Qian Yang, Yajuan Huang, Kewei Li, Yusi Fan, Lan Huang, and Fengfeng Zhou. OCMR: A comprehensive framework for optical chemical molecular recognition. *Comput. Biol. Med.*, 163(C), 2023.

[146] Yifei Wang, Yunrui Li, Lin Liu, Pengyu Hong, and Hao Xu. Advancing drug discovery with enhanced chemical understanding via asymmetric contrastive multimodal learning. *Journal of Chemical Information and Modeling*, 65(13):6547–6557, 2025. PMID: 40548496.

[147] David Weininger. Smiles, a chemical language and information system: 1: Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28:31–36, 1988.

[148] Wes McKinney. Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56 – 61, 2010.

[149] Jason Weston, Samy Bengio, and Nicolas Usunier. Large scale image annotation: Learning to rank with joint word-image embeddings. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, pages 335–349. Springer, 2010.

[150] David S Wishart, Craig Knox, An Chi Guo, Anupriya Shrivastava, Murtaza Hassanali, Paul Stothard, Zhan Chang, and Jon Woolsey. Drugbank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Research*, 36(suppl_1):D901–D906, 2008.

[151] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018.

[152] Xu Xie, Fei Sun, Zhaoyang Liu, Shiwen Wu, Jinyang Gao, Bolin Ding, and Bin Cui. Contrastive learning for sequential recommendation. In *arXiv*, 2021.

[153] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[154] Yahui Xu, Yi Bin, Jiwei Wei, Yang Yang, Guoqing Wang, and Heng Tao Shen. Multi-modal transformer with global-local alignment for composed query image retrieval. *IEEE Transactions on Multimedia*, 25:8346–8357, 2023.

[155] Zhanpeng Xu, Jianhua Li, Zhaopeng Yang, Shiliang Li, and Honglin Li. SwinOCSR: End-to-end optical chemical structure recognition using a Swin transformer. *Journal of Cheminformatics*, 14(1):41, 2022.

[156] Chun-Hsiao Yeh, Cheng-Yao Hong, Yen-Chi Hsu, Tyng-Luh Liu, Yubei Chen, and Yann Lecun. Decoupled contrastive learning. In *ECCV*, 2022.

[157] Sanghyun Yoo, Ohyun Kwon, and Hoshik Lee. Image-to-graph transformers for chemical structure recognition. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3393–3397, 2022.

[158] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. In *Transactions of the Association for Computational Linguistics (TACL)*, volume 2, pages 67–78, 2014.

[159] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6720–6730, 2019.

[160] Rowan Zellers, Jae Sung Park, Xuechen Li, Ari Holtzman, Yonatan Bisk, and Ali Farhadi. Merlot: Multimodal neural script knowledge models. In *Advances in Neural Information Processing Systems*, volume 34, pages 23634–23651, 2021.

[161] Zheni Zeng, Yuan Yao, Zhiyuan Liu, and Maosong Sun. A deep-learning system bridging molecule structure and biomedical text with comprehension comparable to human professionals. *Nature Communications*, 13, 12 2022.

[162] Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. MIRACL: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131, 2023.

[163] Ziyang Zhang, Xuan Shi, Yutong Chen, Ruiqi Guo, Zihan Wang, Yichi Zhang, Ling Liu, Zhiyuan Liu, and Jiliang Tang. Mol-instructions: A large-scale biomolecular instruction dataset for large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

[164] Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. Publaynet: largest dataset ever for document layout analysis. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1015–1022. IEEE, Sep. 2019.

[165] Yihan Zhu, Gang Liu, Eric Inae, and Meng Jiang. Moltextnet: A two-million molecule-text dataset for multimodal molecular learning. *arXiv*, 5 2025.

[166] Mohammadreza Zolfaghari, Yi Zhu, Peter V. Gehler, and Thomas Brox. CrossCLR: Cross-modal contrastive learning for multi-modal video representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1450–1459, October 2021.

Appendices

Appendix A

AI Use Policy

I used Generative AI to refine the language and grammatical errors in my dissertation writeup for certain sections. This might cause some readers to notice an unexpected change in style while reading through the manuscript. Some of the tables as shown are converted to latex from raw excel spreadsheets in the interest of time. All other ideas and the fundamental approach taken are original works of research.

Appendix B

Indexes and Their Metadata

The filtered passages obtained from the text extraction stage and the parsed molecule diagrams from the molecule parsing stage described in the previous Chapter are further processed and linked to generate five unique indexes with keys in each index related to other indexes to make the search and retrieval process efficient downstream. The indexes were stored as Pandas [148] dataframes. These allow for fast saving and loading, as well as being directly ingestible into PyTerrier [89] and RDKit, for searching the indexing.

Figure B.1 shows the outline of the indexes created and how reverse lookup enables lookup for all related passages, IUPAC name or linked diagrams. The base index is the (1) passage index. This index forms the most important part of the search process as all modes of search fall back to this index for reverse lookup. It contains a unique passage id (DocNoP) for each passage followed by the filename of the PDF which the passage belongs to and the page in PDF where it was originally extracted from. It is followed by the actual passage text excerpt including any molecule names that have been identified in its raw extracted form. Finally it contains the bounding-box coordinates (top-left, bottom-right) of the spatial location of where the passage exists in the page. The PDF pages were first converted to images at a resolution of 150 DPI. Given a text query, PyTerrier uses this index to present the top passage match candidates along with their passage ids (DocNoP).

To get molecules present in each of these passages, the (2) Passage to ID index is queried to get all the molecule ids (SMILES ID) associated with each of the passage candidates. These ids are first used to get the actual SMILES strings of the molecules available in the (3) ID to SMILES index. To get the IUPAC names of the molecules, the SMILES ID is now looked up in the (4) ID to IUPAC index.

Passage Index

DocNoP	SMILES ID	PDF	Page	Text	x1	y1	x2	y2
1	1	Voronoi.pdf	23	pyrrolopyrimidine derivative and a pharma...	54	68	72	81

Passage to Diagram Index

DocNoP	SMILES ID	Page	x1	y1	x2	y2
1	1	1255	67	12	87	22

Passage to ID Index

DocNoP	SMILES ID
1	1
1	2

ID to SMILES Index

SMILES ID	SMILES
1	N1C=NC=C2C1=CC=N2

ID to IUPAC Name Index

SMILES ID	Name
1	pyrrolopyrimidine

Figure B.1: (Top to bottom): The base passage index; Canonical SMILES lookup index for every unique SMILES by ID; Passage to Diagram Index for every unique molecule by SMILES; IUPAC/Common Name Lookup for each unique SMILES; Unique Passage to Unique SMILES Index

Parallely, any linked diagrams to the passage is also found by using the (5) Passage to Diagram index. Each passage can be linked to more than one unique molecule and this index accounts for those cases by also storing the SMILES ID for the linked diagram. Finally, after all the reverse lookups are finished, every candidate passage is available to the user with its exact spatial position as well the molecules contained in them with their SMILES and IUPAC name. One or more diagrams linked to the passage are also retrieved and provided to the user. Other modes of search such as SMILES or multi-modal are also similar to the above process except that the initial candidates are first traced back to the actual passage ids (DocNoP).

This way of constructing indexes allow efficiently search backwards starting from a single SMILES string or a keyword or an IUPAC name. Given a SMILES string query, top-K similar candidate SMILES can be found by just using the ID to SMILES index and then traced back to the passage index to find all passages that contain the candidate. It also allows to efficiently store our large index efficiently without the need of repetitive rows in any one index as shown in Figure B.1.

Appendix C

Search Interface

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.

Closed source systems such as SciFinder [41] and Reaxys[1] attempt to mitigate the limitations of chemical standards by allowing users to search for single compounds as text or SMILES [147, 105] – a string based representation for a molecule within patents and publications. 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

[1] https://www.reaxys.com/

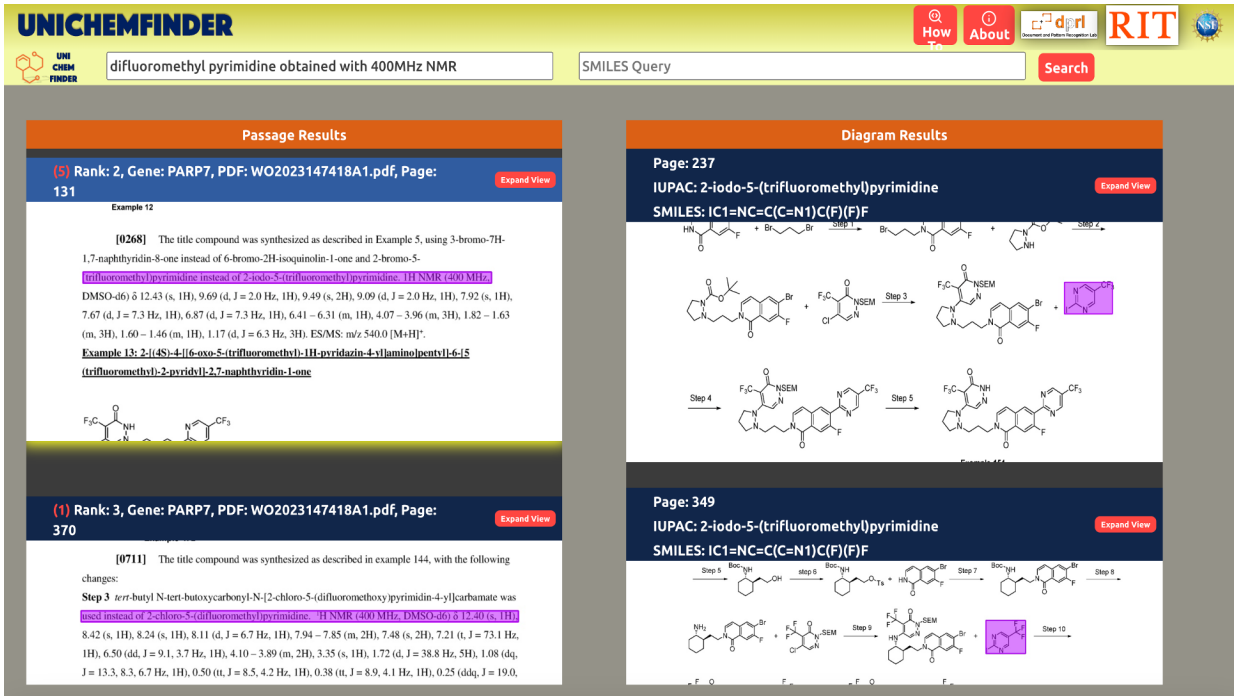

Figure C.1: UniChemFinder page level results for the text-only query "difluoromethyl pyrimidine obtained with 400MHz NMR". Returned passages are shown in the left panel, while diagrams linked to the selected first passage are shown at right in a list. The 'Expand View' buttons allow users to see the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

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 the user interface for UniChemFinder [33] to address limitations in existing CIR systems (see Figure C.1). At a glance, distinguishing features of our system are:

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)

C.0.1 Design Philosophy

Figure C.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 C.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 C.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.

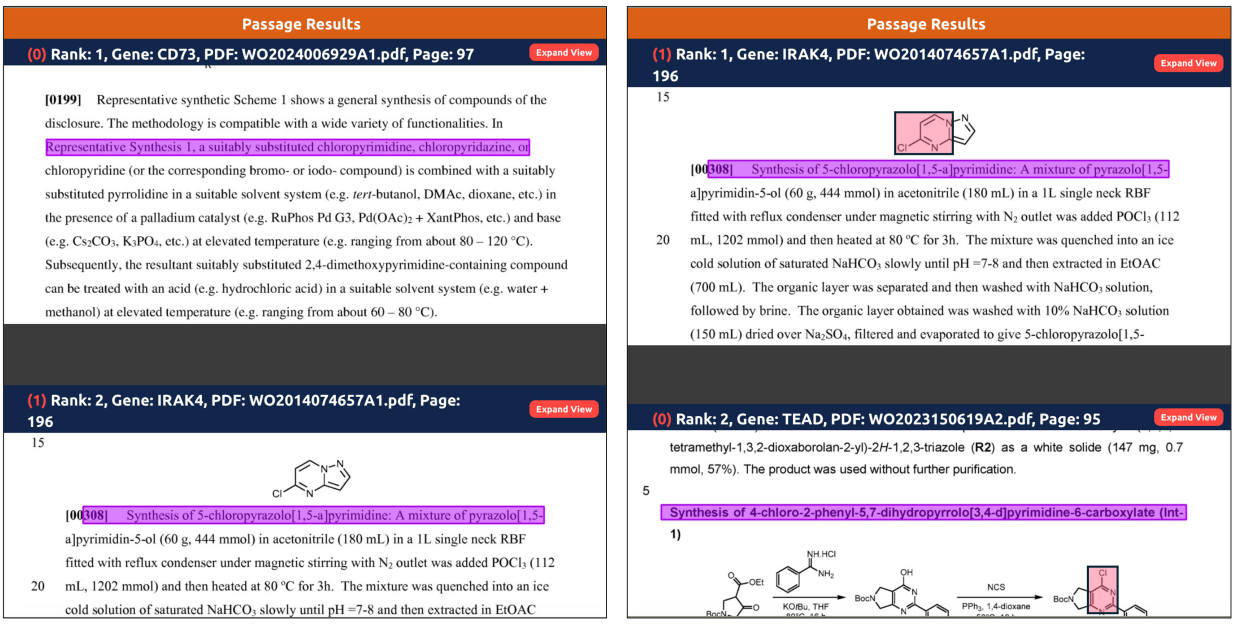

Figure C.2: Left: Search Result from text-only query "synthesis of flourooxetan pyrimidine". Right: Result for multi-modal query combining the text with the SMILES string "ClC1=NC=CC=N1". Right: The manually highlighted substructures are matches for the SMILES part of the query. This demonstrates how 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.

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.

C.0.2 Example Use-Case

Pyrimidine compounds are of 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 C.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, *ClC1=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 C.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.

Appendix D

Test Collection Queries

Table D.1: Multi-modal queries used for expert annotation. Each query consists of a natural language text component and a corresponding SMILES query. Candidate types: Txt = text-only, Diag = diagram-only, T+D = text and diagram.

#	Text Query	SMILES Query	#Scored	Txt	Diag	T+D
1	Show me where this reagent, 1-ethyl-3-(3-dimethylaminopropyl)carbodiimide (ethyldimethylaminopropylcarbodiimide), is utilized to couple an amine with carboxylic acid	`CN1N=C(OC)C(C([N])=O)=C1`	22	10	7	5
2	3-Methoxy-1-methyl-1H-pyrazole-4-carboxylic acid	`CN1C=C(C(OC)=N1)C(O)=O`	12	3	7	2
3	Show me where the substrate 3-(methylthio)-1-phenyl-1H-indazol-6-amine is used in a peptide forming reaction	`CSC1=NN(C2=CC=CC=C2)C3=C1C=CC(N)=C3`	16	4	7	5
4	Show me any characterization or biological assays data of the molecule	`O[C@H](/C=C/C[C@H](C)C[S@]1(NC(C2=CN(C)N=C2OC)=O)=O)[C@H](CC3)[C@@H]3CN4C[C@]5(C(C=CC(Cl)=C6)=C6CCC5)COC7=CC=C(C(N=1)=O)C=C74`	27	12	10	5
5	Show me where an amine undergoes a nucleophilic aromatic substitution with methyl 3-fluoro-4-nitrobenzoate	`O=C(OC)C1=CC(F)=C([N+]([O-])=O)C=C1`	8	4	2	2

Continued on next page

Table D.1 – continued

#	Text Query	SMILES Query	#Scored	Txt	Diag	T+D
6	Show me where this ligand, 4,4'-Di-tert-butyl-2,2'-dipyridyl, is used in a reaction	O=C1N(C2C(N(COCC[Si](C)(C)C)C(CC2)=O)=O)CC3=CC(OC4CNCCC4NC(OC(C)(C)C)=O)=CC=C31	20	10	3	7
7	Show me where the reagent 4,4,5,5-tetramethyl-2-(4,4,5,5-tetramethyl-1,3,2-dioxaborolan-2-yl)-1,3,2-dioxaborolane is used in a reaction	BrC(C=C(P(C)(C)=O)C=C1)=C1NC2=CC=C(S(F)(F)(F)(F)F)C=C2	12	4	3	5
8	Show me where the reagent trifluoroacetic acid is used to remove a Boc protecting group	CC(C)(C)OC(N[C@H](CCC1)CN1C(C(F)(F)F)C2=CC=CC=C2)=O	21	8	7	6
9	Show me where the substrate 2-[(4-chloro-2-fluoro-phenyl)methoxy]-6-(1,2,3,6-tetrahydropyridin-4-yl)pyridine is alkylated	FC(C=C(Cl)C=C1)=C1COC2=CC=CC(C3=CCNCC3)=N2	21	9	6	6
10	Show me where the substrate (3R,4R,SR)-S-((2-(Diethoxyphosphoryl)ethoxy)-methyl)tetrahydrofuran-2,3,4-triyl triacetate is used in a reaction	O=P(OCC)(OCC)CCOC[C@H]1OC(OC(C)=O)[C@H](OC(C)=O)[C@@H]1OC(C)=O	16	8	5	3
11	Buchwald-Hartwig coupling with piperidine	C1(N2CCCCC2)=CC=CC=C1	19	10	6	3

Continued on next page

Table D.1 – continued

#	Text Query	SMILES Query	#Scored	Txt	Diag	T+D
12	Palladium catalyzed reaction with piperidine	C1(N2CCCCC2)=CC=CC=C1	20	8	8	4
13	Aryl halide treated with palladium and piperidine	C1(N2CCCCC2)=CC=CC=C1	19	6	10	3
14	Piperidine treated with a metal catalyst and aryl halide	C1(N2CCCCC2)=CC=CC=C1	18	7	7	4
15	Isolation of aryl amine with palladium	C1(N2CCCCC2)=CC=CC=C1	23	11	7	5
16	Amide coupling with piperidine	CC(N1CCCCC1)=O	20	13	5	2
17	Treatment of a carboxylic acid with piperidine	CC(N1CCCCC1)=O	14	5	6	3
18	Amide coupling with HBTU	CC(N1CCCCC1)=O	26	12	7	7
19	Mixture of piperidine and HBTU	CC(N1CCCCC1)=O	20	8	8	4
20	Mixture of aniline and HBTU	CC(N1CCCCC1)=O	15	4	5	6
21	A heteroarene treated with morpholine	C1(N2CCOCC2)=CC=NC=C1	17	5	8	4
22	A pyridine treated with morpholine	C1(N2CCOCC2)=CC=NC=C1	13	5	6	2
23	A nucleophilic aromatic substitution with pyridine	C1(N2CCOCC2)=CC=NC=C1	22	10	6	6
24	A nucleophilic aromatic substitution with pyridine and morpholine	C1(N2CCOCC2)=CC=NC=C1	18	6	10	2
25	Arylation of morpholine	C1(N2CCOCC2)=CC=NC=C1	14	4	8	2
26	A synthetic sequence with a Suzuki reaction	C1(C2=CC=CC=C2)=CC=CC=C1	25	12	9	4

Continued on next page

Table D.1 – continued

#	Text Query	SMILES Query	#Scored	Txt	Diag	T+D
27	A synthetic sequence with the formation of an aryl-aryl bond	C1(C2=CC=CC=C2)=CC=CC=C1	27	13	11	3
28	Treatment of an aryl halide with an arylboronic acid	C1(C2=CC=CC=C2)=CC=CC=C1	21	10	9	2
29	Treatment of an arylboronic acid with palladium	C1(C2=CC=CC=C2)=CC=CC=C1	18	10	6	2
30	Formation of a biaryl	C1(C2=CC=CC=C2)=CC=CC=C1	27	15	7	5
31	A synthetic sequence with a carbamate protection	O=C(OC(C)(C)C)NC1=CC=CC=C1	23	11	5	7
32	Treatment of an amine with carbamoyl chloride	O=C(OC(C)(C)C)NC1=CC=CC=C1	20	7	2	11
33	NBoc protection	O=C(OC(C)(C)C)NC1=CC=CC=C1	26	15	2	9
34	Aniline protection with FMoc	O=C(OC(C)(C)C)NC1=CC=CC=C1	15	8	1	6
		Total	**655**	**287**	**216**	**152**