MoleculeMiner: Extracting and Linking Molecule Figures with Tabular Metadata

Abhisek Dey[1],[2] , **Nathaniel H. Stanley**[2]
[1]Rochester Institute of Technology
[2]Insitro
ad4529@rit.edu, nate@insitro.com

Abstract

Despite an ongoing shift in automated chemical literature search methods, many are fairly limited in ability to find very specific relevant information about a drawn molecule and its associated property data. We aim to tackle the challenge of converting drawn molecules to a machine readable representation and co-reference any associated molecule data. *MoleculeMiner* is a system where a user can feed in their own patent or paper to obtain each drawn molecule along with any specific metadata (chemical name, chemical reactivity, yield, purity etc.) provided anywhere in the PDF in a tabular format, using an interactive user-friendly environment. We also present *MolScribeV2*, a molecular image parser which improved upon the original MolScribe by introducing pixel-based self attention positional embedding technique. Along with other changes, MolScribeV2 is robust to varied styles of compound drawings commonly found in patents and papers–scanned or born digital. Our extraction and user interactive system can be found at https://github.com/insitro/MoleculeMiner.

1 Introduction

While structured extraction from unstructured documents in the chemical domain is not new, most of these systems have modality and input format limitations. Some recent systems for chemical structure recognition (CSR) include DECIMER [Rajan *et al.*, 2020] that used a CNN based encoder-decoder architecture to directly predict SMILES from a molecule image. MolScribe [Qian *et al.*, 2023], which MoleculeMiner's parsing is based off of, used a transformer based encoder-decoder model but was limited in performance due to being trained on synthetically generated data. MolGrapher [Morin *et al.*, 2023] used a different approach by first using a CNN to predict the atoms and then use a graph neural network (GNN) to predict the graph structure of the molecule. This was then converted to a SMILES representation. ChemScraper [Shah *et al.*, 2024] on the other hand used two different parsing techniques for born-digital (rules) and image-based (neural) to improve accuracy if the PDF has embedded graphical elements for the molecule figures. MoleculeMiner aims to improve upon the existing MolScribe CSR system by adding a novel positional embedding technique and addition of new forms of data augmentation for CSR to be robust to diagrams extracted from patents as well as journals.

Extracting tabular data in the chemical domain has always been a challenge due to the large variations in table styles found. Further, these tables often don't contain the chemical structure, and instead have a text identifier to the compounds. ChemTables [Zhai *et al.*, 2021] tried an approach by trying to classify the type of table first by populating a table type ontology from Reaxys [Reaxys, 2009]. ChemDataExtractor [Swain and Cole, 2016] used a layout classifier to identify tables from other objects such as images and text blocks based on a heirarchy of rules. The segmented tables were then parsed row-wise based on embedded characters in the PDF. Thus, it could only support born-digital documents as their base extraction system was based on PDFBox [The Apache Software Foundation, 2012]. Modern approaches introduce using vision based multi-modal Large Language Models (LLM) for zero-shot extraction of table data given a specific schema through prompt engineering. ChatExtract [Polak and Morgan, 2024] used a hierarchy of prompts to extract chemical data from tables in a JSON format. MoleculeMiner attempts to automate the table extraction and linking using carefully constructed prompts to identify chemically relevant tables.

Co-referencing or entity-linking different modes of information sources from a document in the chemical domain is an equally challenging task due to variations in document generation methods and structure. [Zhang *et al.*, 2023] used a BiLSTM based Named Entity Recognition (NER) model to extract relevant chemical text and a CNN model to perform Optical Character Recognition (OCR) on document tables to link material names in text with associated units and composition data in tables. However their table schema that were filtered had to adhere to a fixed expected schema. ChemSchematicResolver [Beard and Cole, 2020] linked R-group definitions in text to diagrams containing open group substituent. OpenChemIE [Fan *et al.*, 2024] is an end-to-end system that used a table detection method and identifying relevant ones though table headers to replace R-groups in diagrams. Our system can identify links between any form of structured tables and diagrams as long as there is a reference number written in its vicinity.

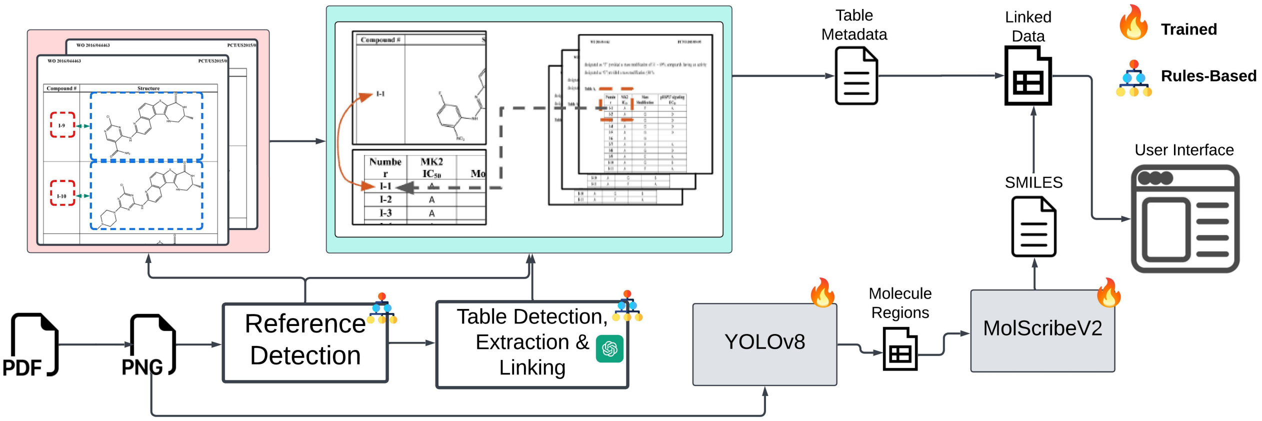

Figure 1: Pipeline Overview. A PDF provided by the user through the interface is processed to localize molecule regions in page using YOLOv8. MolScribeV2 is used to convert these molecule regions into their SMILES representations. In parallel, a reference number for the drawing was identified, if any. These references were finally used to link to data extracted from tables fusing GPT-4o and the associated molecule which is served back to the user as an interactive overlay.

Figure 1 shows the overall pipeline of MoleculeMiner. When a user provides a PDF document (born-digital or scanned), the system runs it through a multi-stage pipeline that (1) Detects, extracts and parses Drawn Molecule Diagrams into a Simplified Molecular Input Line Entry System (SMILES) [Weininger, 1988] string *(Chemical Structure Recognition (CSR))* (2) Detects and extracts all data from tables *(Table Extraction)* (3) Co-references molecule property data from tables and links them to parsed molecules *(Entity Linking)*.

2 System Description

2.1 Molecule Detection and Parsing

The detection and parsing system first identifies the regions where the molecules exists, then parses the individual molecules through a transformer based encoder-decoder system 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 – applicability in molecular downstream tasks, encoding of chemical properties such as chirality and double-bond geometry (stereochemistry) and easy interconversion to other formats such as MOLFile, DeepSMILES [O'Boyle and Dalke, 2018] and InChI [Heller *et al.*, 2015] [Heller, 2014].

Molecule Diagram Detection: PDF pages are processed through YOLOv8 [Jocher *et al.*, 2023] to segment the molecule regions. CLEF-IP2012 [Piroi *et al.*, 2012] [Dey and Zanibbi, 2021] dataset consisting of 1242 pages and 419 pages for train and test respectively containing molecule diagrams. Due to limitation in availability of open source training data for segmentation, we avoid using data intensive options like vision transformer for this stage.

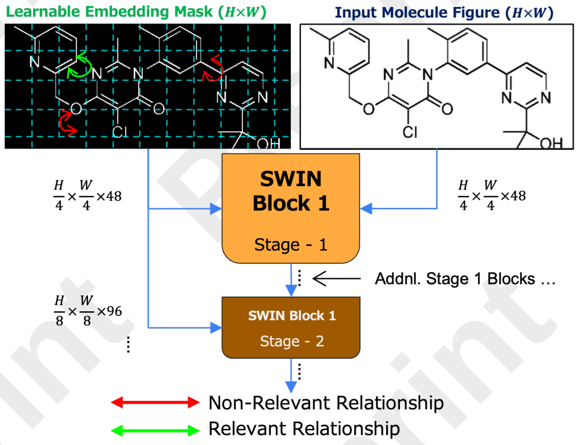
Learnable Embedding Mask (H×W)
Input Molecule Figure (H×W)
$\frac{H}{4}\times\frac{W}{4}\times48$
$\frac{H}{4}\times\frac{W}{4}\times48$
SWIN Block 1
Stage - 1
Addnl. Stage 1 Blocks ...
$\frac{H}{8}\times\frac{W}{8}\times96$
SWIN Block 1
Stage - 2
Non-Relevant Relationship
Relevant Relationship

Figure 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.

Molecule Diagram Parsing: Our parsing model, MolScribeV2, improves upon the original MolScribe [Qian *et al.*, 2023] in three major areas – enhanced positional embeddings, dataset size and special augmentation. Figure 2 shows the additional positional embedding method where at each block of the SWIN transformer encoder [Liu *et al.*, 2021], an additional learnable positional embedding is created from the binary mask of graphics pixels that forces attention values in the transformer layers to distinguish between valid and invalid pixels. Compared to MolScribe which was trained on 1 million PubChem [Kim *et al.*, 2022] molecules, 5 million molecules were used to train MolScribeV2 with data from PubChem, Zinc250k [Akhmetshin *et al.*, 2021], ChemBL [Mendez *et al.*, 2019], MOSES [Polykovskiy *et al.*, 2020]. Finally, existing data augmentation methods were combined with two new techniques to improve performance on real world data. Margin cropping or adding was a method used

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)

Table 1: 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.

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. 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. Table **??** shows the comparison of original version to various improvements on a set of two real-world patent PDFs. 1832 molecules were manually collected along with their ground-truth SMILES. This set was collated with expert annotated SMILES to produce the Page → Location → Molecule → SMILES examples from real world documents. This is the first such collection made to our knowledge and will be made available. We do not report benchmark comparisons due to space constraints and those not reflecting real world extracted molecules.

2.2 Molecule and Table Data Linking

The process involves three main stages. (1) Finding the reference number written in the proximity to each drawn molecule (2) Extracting all tables found in the PDF and filtering only the tables containing a column with reference numbers (3) Merging drawn molecules with their associated table metadata based on common reference numbers.

References to drawn molecules were found using the docTR [Mindee, 2021] OCR suite. After filtering out all words that either were not numeric or alphanumeric or were inside a molecule region, the rest of the words were chosen as candidate sets for attaching to each individual molecule. Proximity-based matching included spatial constraints such as reference numbers can only be to the bottom or right of a molecule, each reference number instance can only be used without replacement and have to fall within certain distance based on the longest side of molecule region. This approach was found to be most consistent across different styles of documents.

Table Detection and Extraction was performed through GPT-4o [et al., 2024] through a series of prompts. The prompts were carefully designed to sequentially extract the table label, table headers and the row metadata. Additional algorithms were designed to join multi-page tables together, process tables with no table headers and filter out columns where the molecule is drawn inside the table itself.

Linking molecules with their associated metadata involved collating reference numbers found near drawn molecules with reference numbers in tables. In our work, we found reference column headers could be named in a different ways and specific keyword list was designed to account for the variations.

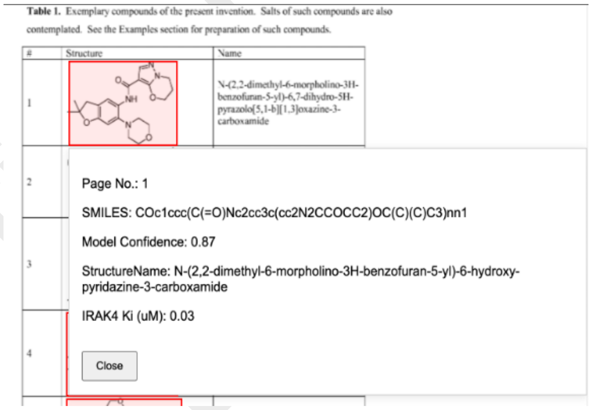

Table 1. Exemplary compounds of the present invention. Salts of such compounds are also contemplated. See the Examples section for preparation of such compounds.

#	Structure	Name
1		N-(2,2-dimethyl-6-morpholino-3H-benzofuran-5-yl)-6,7-dihydro-5H-pyrazolo[5,1-b][1,3]oxazine-3-carboxamide
2		
3		
4		

Page No.: 1

SMILES: COc1ccc(C(=O)Nc2cc3c(cc2N2CCOCC2)OC(C)(C)C3)nn1

Model Confidence: 0.87

StructureName: N-(2,2-dimethyl-6-morpholino-3H-benzofuran-5-yl)-6-hydroxy-pyridazine-3-carboxamide

IRAK4 Ki (uM): 0.03

Figure 3: An excerpt from the user interface showing detected molecules on a PDF page and a clickable pop-up menu showing a specific molecule's extracted and linked properties from anywhere in the PDF through reference matching with tables.

The approach taken in this paper for co-referencing improves previous systems in various ways. Multi-page tables can be successfully joined together, forming a unified coherent table for linking regardless of table schema change between pages. Each molecule can be linked to metadata obtained from more than one table. Any type of metadata including multi-line metadata can be successfully linked without the need of constraining to very specific table schemas.

3 User Interface

To facilitate user adoption of Molecule Miner, we have created a user interface allows users to run analyses with little effort, including those lacking programming experience (see video accompanying paper). The user can upload a PDF of interest and, optionally, their OpenAI API key if they want table extraction as well. Once complete, the user is served an interactive display of their document with clickable red boxes indicating an identified molecule. Upon clicking each red box, the user can see information about the molecule like its SMILES and the model confidence. Figure 3 shows available metadata collated from the entire PDF for a specific molecule displayed to the user. Furthermore, a user can directly download the full PDF result as a CSV file that can be directly used in downstream applications like molecule modeling, property prediction etc.

4 Conclusion

MoleculeMiner uses specialized extraction and co-referencing methods, linking parsed diagrams with its pharmacologically relevant data mentioned in tables. This is enabled by our robust molecule diagram extraction and improved parsing along with the usage of Large Language Models to detect and parse tables. Automated linking of relevant table data with molecules allows important molecule properties to be easily accessible to be used in other downstream tasks for cheminformatics. The webapp exposes this functionality in an easy-to-use interface through which users can get information from both digital and scanned PDFs.

Acknowledgments

We would like to thank our colleagues in DS-ML team at Insitro for their contributions in making this project including Patrick Conrad, Srinivasan Sivanandan, Kent Gorday, Tomasz Danel, and Benson Chen for their thoughts and ideas and infrastructure help for this project.

References

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

[Beard and Cole, 2020] Edward J. Beard and Jacqueline M. Cole. Chemschematicresolver: A toolkit to decode 2d chemical diagrams with labels and r-groups into annotated chemical named entities. *Journal of chemical information and modeling*, 60:2059–2072, 2020.

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

[et al., 2024] Josh Achiam et al. Gpt-4 technical report, 2024.

[Fan *et al.*, 2024] 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.

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

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

[Jocher *et al.*, 2023] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolov8, 2023.

[Kim *et al.*, 2022] 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.

[Liu *et al.*, 2021] 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.

[Mendez *et al.*, 2019] 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.

[Mindee, 2021] Mindee. doctr: Document text recognition. https://github.com/mindee/doctr, 2021.

[Morin *et al.*, 2023] 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.

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

[Piroi *et al.*, 2012] 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.

[Polak and Morgan, 2024] Maciej P. Polak and Dane Morgan. Extracting accurate materials data from research papers with conversational language models and prompt engineering. *Nature Communications*, 15, 12 2024.

[Polykovskiy *et al.*, 2020] 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.

[Qian *et al.*, 2023] 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.

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

[Reaxys, 2009] Reaxys. *Reaxys*. [Frankfurt, Germany] ; [New York, NY] : Elsevier, 2009. System requirements: Sun Microsystems Java Version 5.0 or later (also known as Java Runtime Environment (JRE) Version 1.5.0).

[Shah *et al.*, 2024] 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.

[Swain and Cole, 2016] 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.

[The Apache Software Foundation, 2012] The Apache Software Foundation. Apache pdfbox projekt, 2012.

[Weininger, 1988] 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.

[Zhai *et al.*, 2021] Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16 x 16 words :. *International Conference on Learning Representations*, pages 1–21, 2021.

[Zhang *et al.*, 2023] Rui Zhang, Jiawang Zhang, Qiaochuan Chen, Bing Wang, Yi Liu, Quan Qian, Deng Pan, Jinhua Xia, Yinggang Wang, and Yuexing Han. A literature-mining method of integrating text and table extraction for materials science publications. *Computational Materials Science*, 230, 10 2023.