ChemSIE: From Document Based Records to Machine Actionable Experimental Data

Itamar Wallwater [1] **Yonatan Harnik** [1] **Ari Pakman** [2] **Anat Milo** [1]

[1]*Department of Chemistry, Ben-Gurion University of the Negev, Beer Sheva 84105, Israel* [2]*Department of Industrial Engineering and Management, Ben-Gurion University of the Negev, Beer Sheva 84105, Israel*. Correspondence to: Itamar Wallwater dropides@void.ai.

1. Introduction

Experimental characterization data generated in chemical research often originates in machine readable form, yet is subsequently rendered into text based formats designed primarily for human interpretation. For example, experimental characterization information is frequently collected and transmitted in records ranging from Supplementary Information (SI) and theses to internal reports and patent applications, where molecular representations and quantitative measurements are presented together [1]. Nevertheless, the ever increasing amount of experimental chemical data has created new opportunities for data driven discovery [2], while simultaneously exposing limitations in how such data are recorded, accessed, and reused. In contemporary chemical literature, a growing fraction of this information is reported outside the main manuscript text. These records include experimental observables such as infrared (IR) and nuclear magnetic resonance (NMR) peak lists, which are placed in close proximity to molecular identifiers of the corresponding molecular structures, for example, their 2D image or IUPAC name [3, 4]. Despite their central role in chemical research, these records remain difficult to integrate into automated workflows, as most existing extraction tools prioritize narrative text, reaction descriptions, or reference mining [5, 6, 7, 8, 9]. As a result, large volumes of experimentally relevant data must be manually reprocessed to recover machine actionable representations, even as advances in data driven methods increasingly depend on such structured inputs [10, 11].

2. Substantial section

The pipeline is a user-guided, automated workflow that converts document embedded experimental records into structured, molecule-resolved data. Users initiate the process by selecting PDF documents containing experimental characterization data and molecular depictions via a graphical user interface (GUI), command line interface, or Python API, all of which invoke the same backend processing pipeline. The data is then extracted from the documents and can be validated by the users before the final dataset is constructed. This user-in-the-loop design, illustrated in Figure 1, ensures curation quality while the automated workflow allows scalability throughout the pipeline.

Following document ingestion, the pipeline automatically assembles molecule-resolved experimental records through a sequence of layout-aware processing steps that integrate visual and textual modalities. The core automated phase of the pipeline comprises molecular image extraction, coupled with molecule-centric document segmentation and text parsing, collectively enabling the assembly of molecule-resolved experimental records.

The extraction pipeline was evaluated on a set of 24 PDF files, comprising 11 academic theses and 13 Supplementary Information files drawn from the chemical literature. All documents contained experimental characterization data and molecular structure depictions, with a primary emphasis on organic chemistry. Document length varied substantially, reflecting realistic heterogeneity in formatting and content density. Theses averaged approximately 243 pages, whereas SI files averaged 93 pages. For all documents, the experimental text was manually curated to generate a benchmark ground truth reflecting the reported content, without independently validating the chemical correctness of the original records. A total of 792 molecular images were included in this analysis. For each characterization modality (IR, NMR, MS, and TLC), both ground truth and extracted text were parsed into peak lists. Solvent information, annotations, and descriptive qualifiers were intentionally excluded, focusing the evaluation on recovery of quantitative experimental observables. The accuracy of the extracted records was evaluated against the ground truth by computing precision, recall, and F1 scores summarized in Table **??**. Across the benchmark, the pipeline achieved consistently high extraction performance across both SI files and theses, despite substantial differences in document length and formatting. Observed errors primarily arose from atypical formatting or nonstandard separators, which often led to extra unwanted text.

2.1 Related work

As ChemSIE broadens the landscape of automated experimental data extraction, it is important to contextualize its design relative to existing chemical text-mining frameworks. Most established tools have been developed with the primary goal of extracting information from the main body of scientific manuscripts, where experimental details are embedded within narrative prose. In contrast, a substantial fraction of experimentally relevant data is recorded in semi-structured formats such as Supplementary Information, theses, or internal experimental records, where molecular depictions and characterization data are presented together but are not explicitly encoded for automated reuse. This distinction

in document structure has important implications for how experimental data can be systematically extracted, associated, and integrated into downstream workflows.

ChemDataExtractor is among the earliest and most widely adopted platforms for automated chemical text mining and has demonstrated strong performance in extracting experimental information from narrative text [12, 13]. Its design is well suited to the main body of articles, where chemical entities, reactions, and measurements are described within continuous prose. However, when applied to documents in which experimental data are organized around molecule-specific sections often accompanied by molecular images, ChemDataExtractor does not natively impose molecule-level structure on the extracted output. As a result, experimental records are typically returned without explicit linkage to molecular depictions, requiring additional user defined preprocessing to recover molecule-resolved datasets. While highly effective for reference extraction and text centric analyses, this approach leaves a large class of experimentally rich, image associated records underutilized.

Recent advances in large language model based systems, such as ChemCrow, have further expanded the recall and flexibility of chemical information extraction [14]. These tools excel at reasoning over short text passages and integrating heterogeneous sources of chemical knowledge. Nevertheless, their reliance on text-only inputs and practical constraints on input length necessitate document segmentation or preprocessing for larger corpora, which can complicate automated, large scale data extraction. Moreover, like earlier text mining frameworks, these approaches do not explicitly address the joint extraction and association of molecular images with experimental data, limiting their applicability to molecule-resolved dataset construction. Related efforts such as OpenChemIE [15] provides a powerful framework for standardized chemical information extraction, but, like most text centric approaches, it primarily targets structured representations within the main article body. While these platforms represent important advances in scalable chemical information extraction, they are likewise optimized for main text analysis. Consequently, experimental data recorded outside of the narrative article body remain difficult to integrate into structured, machine readable formats.

Other emerging platforms, including OdanChem, demonstrate promising performance in experimental data extraction [16]. However, their presentation as closed or opaque systems limits transparency into internal decision processes and restricts extensibility within automated data pipelines. For workflows focused on dataset construction, benchmarking, or downstream machine learning, this lack of inspectability and modularity can present a significant barrier.

Rather than seeking to replace existing text centric extraction frameworks, ChemSIE is designed to complement them by addressing a class of experimentally rich records that fall outside the primary scope of current tools. By producing structured, interoperable, molecule-resolved experimental datasets with explicit linkage to source text and imagery, the workflow facilitates data reuse, incremental database growth, and downstream integration with data driven discovery pipelines. In doing so, ChemSIE contributes to the broader goal of transforming document embedded experimental data into machine actionable, FAIR aligned chemical datasets.

Looking forward, the increasing inclusion of explicit molecular identifiers in experimental reports presents an opportunity to further strengthen molecule-resolved data extraction. Rather than obviating the need for document-level integration, such identifiers can be incorporated as an additional modality within frameworks like ChemSIE, improving robustness while preserving explicit links between experimental observables and their documentary context. In this sense, ChemSIE provides a foundation for integrating past, present, and future experimental records into interoperable datasets, even as reporting conventions continue to evolve.

2.2 Figures and tables

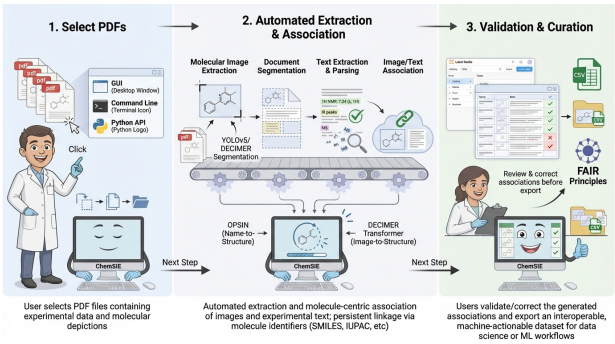

Fig. 1: Overview of the ChemSIE pipeline for constructing molecule-resolved experimental datasets from document based sources. Users initiate processing by selecting PDF documents containing experimental text and molecular depictions. The pipeline then performs automated layout aware processing, including molecular image detection, molecule-centric document segmentation, experimental text parsing, and cross modal association, to assemble preliminary molecule-resolved experimental records. Users can subsequently validate molecule identities and experimental associations before final dataset construction. The resulting curated dataset is machine actionable, interoperable, and ready for downstream analysis

Acknowledgments

A. Milo is supported by the Israel Science Foundation (Grant No. 2252/21). A. Pakman is supported by the Israel Science Foundation (Grant No. 1138/23). I. Wallwater acknowledges the Kreitman Graduate

Table 1: Major data of the orbits of Venus, Earth, Mars and Jupiter.

Subset	Number of samples	Precision	Recall	F1	Name score	SMILES score
^{1}H NMR	551	98.4	98.4	98.4	-	-
^{13}C NMR	564	92.4	97.9	94.4	-	-
IR	430	97.4	97.8	97.0	-	-
MS	545	94.4	95.3	93.6	-	-
Overall	2034	95.6	97.3	95.8	76.4	87.5

School for her Ph.D. funding. I. Wallwater also gratefully acknowledges Michael Schneider for the name of the extractor

References

[1] R.M. Silverstein, F.X. Webster, and D. Kiemle. *Spectrometric Identification of Organic Compounds, 7th Edition*. Wiley, 2005.

[2] Robert Pollice, Gabriel dos Passos Gomes, Matteo Aldeghi, Riley J. Hickman, Mario Krenn, Cyrille Lavigne, Michael Lindner-D'Addario, AkshatKumar Nigam, Cher Tian Ser, Zhenpeng Yao, and Alán Aspuru-Guzik. Data-driven strategies for accelerated materials design. *Accounts of Chemical Research*, 54(4):849–860, 2021. PMID: 33528245.

[3] Mihai Pop and Steven L Salzberg. Use and misuse of supplementary material in science publications. *BMC bioinformatics*, 16(1):237, 2015.

[4] Mathias Christmann. What I Learned from Analyzing Accurate Mass Data of 3000 Supporting Information Files. *Organic Letters*, 27(1):4–7, 2025. PMID: 39700618.

[5] 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(14):5521–5534, 2024.

[6] Daniel Mark Lowe. *Extraction of chemical structures and reactions from the literature*. PhD thesis, University of Cambridge, 2012.

[7] C. Ramakrishnan, A. Patnia, E. Hovy, et al. Layout-aware text extraction from full-text pdf of scientific articles. *Source Code for Biology and Medicine*, 7(1):7, 2012.

[8] A. Tharatipyakul, S. Numnark, D. Wichadakul, and S. Ingsriswang. ChemEx: information extraction system for chemical data curation. *BMC Bioinformatics*, 13 Suppl 17(Suppl 17):S9, 2012.

[9] Miao Zhu and Jacqueline M. Cole. PDFDataExtractor: A tool for reading scientific text and interpreting metadata from the typeset literature in the portable document format. *Journal of Chemical Information and Modeling*, 62(7):1633–1643, 2022.

[10] Sally Bloodworth, Cerys Willoughby, and Simon J. Coles. Data accessibility in the chemical sciences: an analysis of recent practice in organic chemistry journals. *Beilstein Journal of Organic Chemistry*, 21:864–876, 2025.

[11] M. Wilkinson, M. Dumontier, and I. et al. Aalbersberg. The fair guiding principles for scientific data management and stewardship. *Sci Data 3*, 2016.

[12] 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(10):1894–1904, 2016.

[13] Shu Huang and Jacqueline M Cole. A database of battery materials auto-generated using ChemDataExtractor. *Scientific Data*, 7(1):260, 2020.

[14] Andres M Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Chemcrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376*, 2023.

[15] OpenChemIE Contributors. Openchemie. `https://github.com/your-org/OpenChemIE`, 2024. Accessed: 2025-06-06.

[16] Oleg Afanasyev, Fedor Kliuev, Alexandra Balalaeva, and Denis Chusov. Rethinking reaction discovery: From serendipity to data-driven chemistry. 2025.