

MarkushGrapher: Joint Visual and Textual Recognition of Markush Structures

Lucas Morin^{1,2} Valéry Weber¹ Ahmed Nassar¹ Gerhard Ingmar Meijer¹
Luc Van Gool^{2,3} Yawei Li² Peter Staar¹
¹IBM Research ²ETH Zurich ³INSAIT

{lum, vwe, ahn, inm, taa}@zurich.ibm.com {yawei.li, vangool}@vision.ee.ethz.ch

Abstract

The automated analysis of chemical literature holds promise to accelerate discovery in fields such as material science and drug development. In particular, search capabilities for chemical structures and Markush structures (chemical structure templates) within patent documents are valuable, e.g., for prior-art search. Advancements have been made in the automatic extraction of chemical structures from text and images, yet the Markush structures remain largely unexplored due to their complex multi-modal nature. In this work, we present MarkushGrapher, a multi-modal approach for recognizing Markush structures in documents. Our method jointly encodes text, image, and layout information through a Vision-Text-Layout encoder and an Optical Chemical Structure Recognition vision encoder. These representations are merged and used to autoregressively generate a sequential graph representation of the Markush structure along with a table defining its variable groups. To overcome the lack of real-world training data, we propose a synthetic data generation pipeline that produces a wide range of realistic Markush structures. Additionally, we present M2S, the first annotated benchmark of real-world Markush structures, to advance research on this challenging task. Extensive experiments demonstrate that our approach outperforms state-of-the-art chemistry-specific and general-purpose vision-language models in most evaluation settings. Code, models, and datasets are available ¹.

1. Introduction

Document understanding plays a critical role in accelerating discovery in chemistry. For example, databases [32] can be built to offer a unified view of the current knowledge in documents. Such databases can substantially facilitate and accelerate Research & Development in fields ranging from drug discovery to material science [37]. In documents,

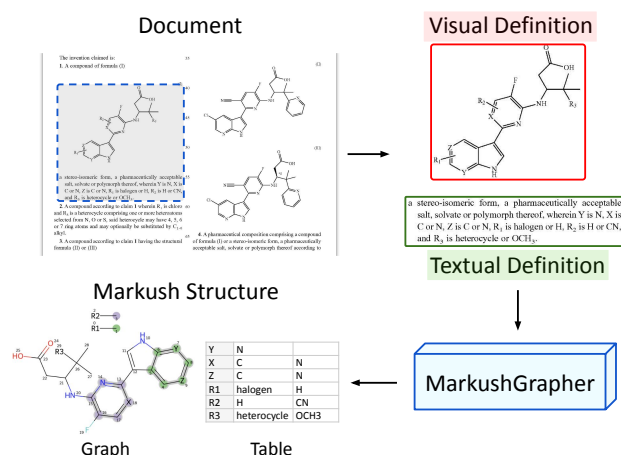


Figure 1. **MarkushGrapher** extracts Markush structures from documents using their visual and textual definitions.

molecules can be referred to using textual names or depictions of their molecular structures. Over the past years, substantial advances have been made to extract molecules from text using chemical name entity recognition [26, 43, 56], as well as, from images through chemical structure image segmentation [40, 45, 57] and chemical structure recognition [31, 38, 41]. More recently, advances in multi-modal document understanding have opened up opportunities to extract chemical structure templates, the so-called Markush structures [42]. These widely used representations are compact descriptions of families of related molecules. A Markush structure consists of a combination of an image and a text element: the image defines the Markush backbone, which contains atoms, bonds, and variable regions, while the accompanying text element specifies the molecular substituents that can replace those variable regions. The variable regions can be variable groups (commonly referred to as R-groups), frequency variation indicators, and positional variation indicators [12]. A visualization of these variable regions is provided in the suppl. materials. These structures are particularly valuable for prior-art search, freedom-to-operate assessment, or landscape analysis in patent documents [42]. Yet, only a very few databases containing

¹<https://github.com/DS4SD/MarkushGrapher>

Markush structures exist (CAS MARPAT [13], CAS DW-PIM [15], and WIPO Patentscope), all of which are proprietary and manually created. Multi-modal Markush Structure Recognition (MMSR) is a critical task towards automatically creating and scaling Markush structure databases.

MMSR presents several substantial challenges that currently limit the effectiveness of automatic recognition methods. First, the immense diversity of Markush backbones, resulting from the combinatorial explosion of atom-and-bond associations, makes good generalization particularly difficult for data-driven deep learning models. Second, Markush backbones can be depicted in a wide range of styles and conventions. Third, Markush textual descriptions employ various formats for representing substituents, such as non-structured names, molecular string identifiers, and chemical formulas. In more complex cases, the descriptions may contain interdependent variable definitions. Finally, the lack of real-world training datasets adds another layer of complexity to applying deep learning methods in this field.

Initial approaches focused on subparts of the MMSR task. Among them, some methods can recognize a subset of Markush backbone images [10, 38, 41]. From the cropped image, they apply a deep network to directly output a string of characters identifying the Markush backbone, in the form of a Simplified Molecular Input Line Entry System (SMILES) [53] or a sequential graph representation. These models extend the task of recognizing molecule images, known as Optical Chemical Structure Recognition (OCSR), by generalizing the recognition of abbreviations to variable groups. However, variable groups are only one specific feature of Markush structure, and these methods do not handle the full complexity of Markush backbones, such as frequency variation indicators. Other works are designed to analyze variable groups definitions from textual descriptions only. Such methods are formulated as a classification task, where labels, and definitions in the textual description are classified. These text-only and image-only methods are limited as they do not exploit the dependencies between variables' textual definitions and their context in the Markush backbone.

Recently, some toolkits have combined textual and visual information for chemical document understanding [14, 49]. However, they still rely on separate models to handle both modalities. At the same time, document understanding has seen the emergence of multi-modal models [20, 46]. They perform a wide range of tasks on documents, using a unified architecture which takes visual and textual inputs, and outputs images and texts as well. Such document understanding models were trained on documents in chemistry [52], and more specifically, to answer if a query molecule is covered by the Markush structures in a document [9]. In this scenario, the model must implicitly recognize multi-modal Markush structures. However, using this method to search

through large sets of documents is not practical, because it would require to run the model on every document for each user query.

In this work, we introduce MarkushGrapher for Multi-modal Markush Structure Recognition, illustrated in Figure 1. The encoder-decoder model takes as input an image of a Markush structure and the OCR cells of all text depicted in the image, and outputs a text sequence representing the structure. This output sequence contains two parts, one represents the graph of the Markush backbone image, and one represents a table of substituents which instantiates the variables in the backbone. The image and OCR cells are jointly encoded using two encoders, a Vision-Text-Layout (VTL) encoder and an Optical Chemical Structure Recognition (OCSR) encoder. These encodings are then concatenated and used by a text decoder to autoregressively generate a Markush sequential representation.

To further aid the research in the community, we introduce M2S, a benchmark dataset of annotated real Markush structures. We also release our pipeline for generating synthetic Markush structures used to train MarkushGrapher. Comprehensive experiments are performed on three benchmarks: MarkushGrapher-Synthetic, M2S and USPTO-Markush. Our model outperforms general-purpose as well as chemistry-specific document understanding models. In summary:

- We develop MarkushGrapher, which combines a VTL encoder and an OCSR encoder to recognize multi-modal Markush structures. Our method achieves state-of-the-art performance for joint recognition of visual and textual definitions of Markush structures.
- We introduce M2S, a benchmark of manually-annotated multi-modal Markush structures from patent documents.
- We build a synthetic data generation pipeline to generate a wide variety of images and accompanying text descriptions of Markush structures.

2. Related Work

Markush Structure Recognition. Existing work on Markush structure recognition typically treats vision and text understanding as separate tasks. On one hand, some models focus exclusively on visual understanding of Markush structures. These recognition methods are often adapted from OCSR techniques [34, 35], which targets standard molecular structures. They commonly employ a vision encoder paired with an autoregressive text decoder, as seen in DECIMER [41], MolScribe [38], MolNextTR [10], and MPOCSR [28]. By only covering variable groups represented as abbreviations, these methods omit most Markush structure features. While Image2SMILES [22] incorporates Markush images with position variation indicators, it remains limited in scope. In contrast, our method tackles all Markush structure features: variable groups, fre-

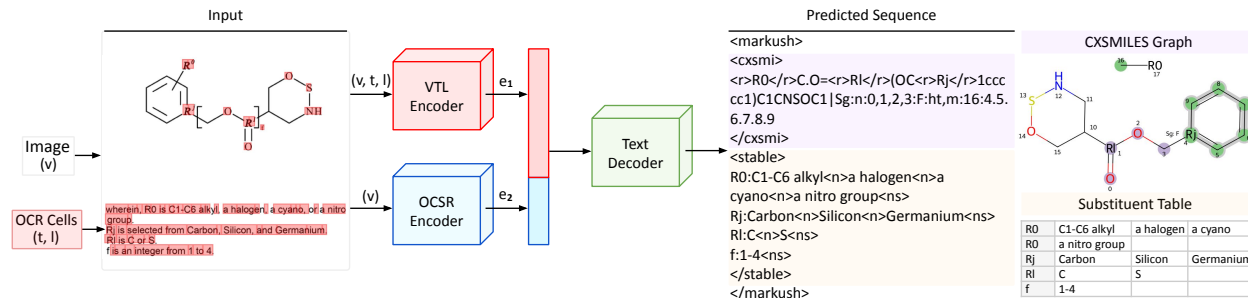


Figure 2. **Markush structure recognition architecture.** MarkushGrapher jointly encodes the input image and its text with a VTL encoder (blue) and an OCSR encoder (red). The VTL output (e_1) and the OCSR output (e_2) are concatenated. Finally, this joint encoding is processed with a text decoder to predict a sequential representation of the Markush backbone (purple) and its substituent table (orange).

quency and positional variation indicators. On the other hand, some approaches exclusively target the recognition of the textual Markush structure component. This task can be formulated as a classification task, and approached, for example, using BioBERT [25] or hand-written grammars [29]. Additionally, several toolkits aim to combine text-only and image-only models for document understanding in chemistry, including CliDE [49], for chemical structure recognition, and OpenChemIE [14] for reaction recognition. Specifically, Beard and Cole [7], Haupt [16], and Jie, *et al.*, [51] proposed toolkits for reconstructing multi-modal Markush structures from text-only and image-only models. However, for MMSR, it is preferred to jointly analyze visual and textual modalities to effectively exploit the dependencies between textual variable definitions and their context in the Markush backbone (see Section 5.6). While Jie, *et al.*, generate a structural library from Markush structures, our method extracts a compact representation. This is crucial for scalability as, for example, Wagner, *et al.*, estimated that a single patent EP0810209B1 already contains about 5×10^{16} molecules [50].

Document Understanding. Document understanding involves extracting a document’s content and structure from a page image input [5, 30]. Recent advancements integrate textual and visual information through unified models, addressing multiple tasks including document classification, layout segmentation, question answering, information extraction, key-value extraction, table structure recognition, and chart question answering [19, 20, 23, 33, 52, 55]. Certain models, such as UDOP [46], rely on an external OCR input and introduce a technique for merging these OCR text tokens together with the image patches containing them. More specifically, initial progress has been made in extending general document understanding models for the recognition of Markush structures. Uni-SMART [9] determines whether a query molecule can be found in a document, directly as a molecule, or covered by one of the Markush structures in the document. However, this approach does not enable the explicit extraction of Markush structures, rendering it impractical for large-scale applications.

3. MarkushGrapher

We introduce MarkushGrapher, a Vision-Text-Layout transformer for MMSR. The model architecture is illustrated in Figure 2. First, the input is encoded using a unified Vision-Text-Layout (VTL) encoder and a Optical Chemical Structure Recognition (OCSR) vision encoder. Second, a text decoder autoregressively predicts from this encoding an optimized Markush sequential representation. Finally, this output is parsed to resolve the graph of the Markush structure backbone and its table of variable groups.

3.1. Architecture

MarkushGrapher takes as input three components: an image, the text content in the image, and the positions of this text provided by bounding boxes. For MMSR, the text, visual, and layout modalities are interdependent. On the Markush backbone, text cells represent the atoms, abbreviations, and variable groups, which are positioned via the bounding boxes, while the image defines the bonds and other non-textual objects. Both are needed together to correctly predict structures. Additionally, the text definition of variable groups gives prior information to the neighboring regions of these groups within the image (see Section 5.6). Besides that, relying on an external OCR, instead of training MarkushGrapher with image-only input has multiple advantages. First, an OCR model can be trained on a wide variety of datasets with fine-grained supervision making it highly robust with various typography. Second, using an OCR is needed to recognize complex abbreviations [31] whose diversity can not be fully covered in training sets. Third, recent work on OCSR shows that relying on localized image features greatly reduces the required number of training samples and improves performance [18, 38]. By providing atom positions as input, the model can easily identify such local features. The three input modalities - image, text and position - are then encoded using two separate encoders.

On the one hand, a Vision-Text-Layout encoder jointly encodes the three modalities following the UDOP’s ap-

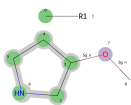
CXSMILES Optimization	Visualization
(1) C[*].C1C(CCN1)OC \$;R1;;;;;;;;;\$;m:0:2.3.4.5.6,Sg:n:7:n:ht	
(2) C<R1</R>.C1C(CCN1)OC m:0:2.3.4.5.6,Sg:n:7:n:ht	
(3) <chem_29><i>0</i></R>.R_1</R><i>1</i></chem_29><i>2</i></chem_17><chem_29><i>3</i></chem_13><chem_29><i>4</i></chem_29><i>5</i></chem_33><i>6</i></chem_17><chem_14><chem_34><i>7</i></chem_29><i>8</i></m:0:2.3.4.5.6,Sg:n:7:n:ht	

Figure 3. **Optimized CXSMILES format.** The figure presents the steps of the CXSMILES optimization. The CXSMILES (1) is first compacted by moving variable groups in the SMILES sequence and removing unnecessary characters (2). Then, the indices of atoms are appended after each atom (between $\langle i \rangle$ and $\langle /i \rangle$ tokens) and the sequence is encoded using a specific vocabulary for atoms and bonds ($\langle chem \rangle$ tokens) (3).

proach [46]. Let e_1 be the output encoding of the VTL encoder,

$$e_1(v, t, l) = \text{VTL}(v, t, l). \quad (1)$$

In this setup, the image (v) is first divided into patches. Image patches and text (t) tokens are independently projected to obtain visual and textual embeddings. Then, the layout (l) information is used to merge both. For each text cell, the embedding of the image patch containing its center is selected. Then, the corresponding visual and textual embeddings are summed. Next, a sequence is formed by appending the joint visual-textual embeddings, followed by the remaining visual embeddings that do not overlap with any text region. The resulting sequence is used as input to a standard transformer encoder. More details on the joint text-vision-layout encoding technique can be found in the UDOP publication [46].

On the other hand, the image is encoded by an OCSR vision encoder. Let e_2 be its output encoding,

$$e_2(v) = \text{OCSR}(v). \quad (2)$$

We use a pre-trained vision encoder implemented as an OCSR transformer. The image is encoded and then projected using a Multi-Layer Perceptron (MLP). Because of the similarities between the Markush structures and chemical structure images, this approach benefits from the robust features learned by OCSR models trained on larger training sets and on real data. This projected OCSR encoding is finally concatenated with the VTL output,

$$e(v, t, l) = e_1(v, t, l) \oplus \text{MLP}(e_2(v)). \quad (3)$$

From this joint encoding, a text decoder iteratively predicts a sequence representing a Markush structure. More details in Section 5.1.

3.2. Optimized Markush Structure Representation

Next, we describe the Markush structure representation in MarkushGrapher, shown in Figure 3. Markush structure backbones are commonly represented using a string identifier named Chemaxon Extended SMILES (CXSMILES)

[3]. The CXSMILES format extends the standard SMILES format, used to store atoms and bonds of molecular structures. CXSMILES strings begin with a standard SMILES, and append an extension block, which allows to represent variable groups, position variation indicators (denoted as ‘m’ sections in the CXSMILES), and frequency variation indicators (denoted ‘Sg’ section in the CXSMILES). Substituents for each variable group can also be stored. Illustrations are provided in the suppl. materials.

To facilitate the auto-regressive prediction of the CXSMILES, we optimized its format (see Fig. 2), making it more compact. Uninformative characters (spaces and ending ‘|’) are removed. To shrink the required attention window, we move the variable groups from the extension section to the SMILES by representing them as special atoms. Also, we add to each atom its SMILES index. As the extension section uses the indices to identify atoms, this extra information explicitly links atoms in the SMILES and the extension section. Moreover, we arrange the atom indices in the ‘m’ sections in ascending order to maintain a consistent format. Finally, we use a specific vocabulary for atoms and bonds rather than encoding them as standard characters. Instead of storing substituents directly in the CXSMILES a substituent table is appended. It enables to store free-text substituents and to compact the definition by grouping together variable groups that have the same substituents, as well as compressing lists of integers defining frequency variation variables.

4. Datasets

4.1. Synthetic Training Data Generation

To the best of our knowledge, no training dataset of annotated Markush structures is available. Thus, we develop a synthetic generation pipeline, described in Figure 4, to generate a broad diversity of visual and textual Markush representations. Generating synthetic Markush structures is a challenge as it requires aligning multi-modal data [51].

Our dataset is created using molecules represented by SMILES selected from the PubChem database [24]. To increase the variety of molecular structures, we sample SMILES based on substructure diversity. Using the RDKit [27] library, these SMILES are augmented to create artificial CXSMILES. (Cf. suppl. material for details.)

Next, we render these CXSMILES using the Chemistry Development Kit (CDK) [54]. Drawing parameters are randomly selected to increase diversity. These include font, bond width, spacing of bonds and atoms, index display as subscripts or superscripts, explicit or implicit carbon display, aromatic cycles with circles, and atom number display. To extract the OCR cells from the synthetic images, we develop a parser for the SVG images generated with CDK. It yields the bounding boxes of text in the image. These

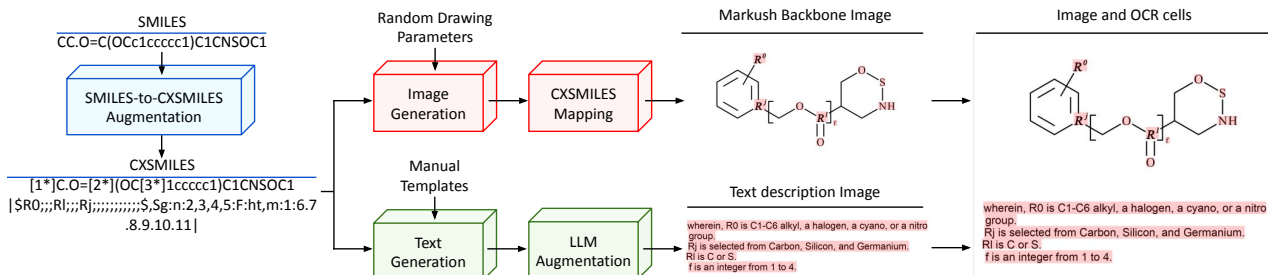


Figure 4. **Synthetic training data generation.** The figure presents the pipeline for generating synthetic training samples. First, a molecule is sampled from PubChem and augmented to create a CXSMILES. Second, the CXSMILES is used to jointly generate a image of the Markush backbone and its OCR cells (red), and generate an image of a text description and its OCR cells (green). Finally, images are collated to create a training sample.

boxes are mapped to their corresponding text, by aligning the SVG file with an associated MOL file [11], which contains the molecular structures.

Furthermore, we generate a text description defining variable groups depicted in the image. For this description to be chemically correct and compatible with the CXSMILES, valence constraints must be respected. The descriptions are generated using multiple templates and mappings. These templates describe the structure of the sentence and are manually extracted from patent documents. They consist of sentences defining variable groups, lists of variable groups, or frequency variation labels. The sentences contain lists of variable groups, lists of substituents, and lists of integers, which are also defined using multiple templates. Additionally, some templates are used to append additional information at the beginning or the end of the description, at the beginning or the end of each item definition, or to add noise information to the description. Then, these templates are instantiated based on mappings, which are lists of manually selected substituents from patent documents, common abbreviations represented by SMILES or names, molecular substructures represented by SMILES or names, and atoms represented by chemical symbols or names. Together with the description, we generate the substituent table.

Finally, we use a LLM to paraphrase a fraction of these descriptions. The substituents table is provided as context to avoid changing any semantic information in the description, but only the writing style. This wide variety of generated descriptions helps generalization. Details on the data generation pipeline are available in the suppl. material.

4.2. The M2S Dataset

There is no public benchmark of annotated real data. Collecting annotations is tedious, hence Markush structure databases usually restrict their textual annotations to a set of keywords [6]. In our case, we chose that the substituents table can contain free-text to cover a large variety of cases.

We introduce M2S, a benchmark of annotated Markush

structures from patent documents. The images are crops of Markush structure backbone images and their textual descriptions. It contains 103 images selected from patents published by the US Patent and Trademark Office (USPTO), European Patent Office (EPO) and World Intellectual Property Organization (WIPO). They are manually annotated with Markush structure backbones, stored as CXSMILES, OCR cells, and substituent tables. The images have a selection bias towards Markush structures in the claims, because these are the most relevant for patents. The Markush structure backbones were annotated on Chemaxon Marvin JS [1], and the OCR cells were annotated using LabelStudio [48]. See suppl. material for further details.

5. Experiments

This section describes experiments with multiple Multimodal Markush Structure Recognition benchmarks.

5.1. Implementation Details

The Vision-Text-Layout encoder and text decoder use a T5-large encoder-decoder architecture [39]. The OCSR encoder is the vision encoder of MolScribe [38]. It is frozen during training. Overall, the model has 831M parameters (744M trainable). It is trained on 210,000 synthetic images, presented in Section 4.1. We train for 10 epochs on a NVIDIA H100 GPU using ADAM with a learning rate of $5e-4$, 100 warmup steps, a batch size of 10 and a weight decay of $1e-3$. The LLM used to augment 10 percent of synthetic text descriptions is Mistral-7B-Instruct-v0.3 [21]. (Cf. suppl. material for details.)

5.2. Evaluation Datasets and Metrics

Datasets. For a comparison with SOTA, our method is evaluated on the benchmarks MarkushGrapher-Synthetic, M2S and USPTO-Markush. We also use a subset of the Sci-Assess benchmark [8] for qualitative evaluation. M2S is described in Section 4.2. MarkushGrapher-Synthetic is a set of 1000 synthetic Markush structures generated using

Table 1. **Comparison of our method with existing MMSR models.** Evaluation on synthetic (MarkushGrapher-Synthetic) and real data (M2S, USPTO-Markush) benchmarks. Exact match (EM) evaluates the percentage of perfect predictions. Tanimoto score (T) and F1-score (F1) evaluate the similarity between the prediction and ground-truth with a percentage between 0 (most dissimilar) and 100 (most similar).

Methods	MarkushGrapher-Synthetic (1000)					M2S (103)					USPTO-Markush (74)	
	CXSMILES		Table		Markush	CXSMILES		Table		Markush	CXSMILES	
	EM	T	EM	F1	EM	EM	T	EM	F1	EM	EM	T
Image-only												
DECIMER [41]	7	35	-	-	-	3	25	-	-	-	7	40
MolScribe [38]	11	47	-	-	-	21	73	-	-	-	7	97
Multi-modal												
Pixtral-12B-2409 [4]	0	3	9	35	0	0	3	7	27	0	-	-
Llama-3.2-11B-Vision-Instruct	0	2	3	7	0	0	1	1	11	0	-	-
GPT-4o [36]	0	6	10	46	0	4	11	19	49	0	-	-
Uni-SMART [9]	-	-	-	-	-	0	46	0	1	0	-	-
MarkushGrapher (Ours)	65	96	84	96	57	38	76	29	65	10	32	74

the pipeline described in Section 4.1. The images are sampled such that overall, each Markush features (R-groups, 'm' and 'Sg' sections) is represented evenly. SciAssess [8] is a benchmark to assess LLMs in scientific literature analysis, covering topics such as biology, chemistry, material science, and medicine. In the chemistry domain, it evaluates if a molecule is contained in a document either, as a molecular structure image, or as a part of one of its Markush structures. The set contains 50 question-answer pairs drawn from 14 documents. As this task is not precisely aligned with our setup, we manually retrieve Markush structure backbone images from these documents and use them for qualitative evaluation. USPTO-Markush is a dataset containing 75 Markush structure backbone images from USPTO patent documents [2]. Its images are annotated with OCR cells and CXSMILES and a large proportion includes indicators for positional and frequency variations.

Metrics. We introduce metrics to evaluate MMSR in two tasks: image recognition and substituent table recognition. For image recognition, we rely on CXSMILES exact match and Tanimoto similarity score [47]. The CXSMILES exact match measures the percentage of perfectly recognized CXSMILES. A match is exact if two conditions are met: (1) disregarding Markush features, the predicted SMILES matches the ground truth according to InChIKey [17] equality, and (2) variable groups, as well as the 'm' and 'Sg' sections, are correctly represented. As to the Tanimoto score, it measures the structural similarity between two chemical compounds. In our case, we remove Markush features, encode structures using the RDKit DayLight fingerprint [44], and compute a similarity score by comparing the bit vectors of these fingerprints using the Tanimoto coefficient. To compute the CXSMILES exact match and Tanimoto score, the stereo-chemistry is ignored. For substituent table recognition, we measure the exact match and the F1-score similarity. The exact match measures the proportion of perfectly recognized tables, *i.e.*, all variable groups and substituents are correct. To compute the F1-score, we first determine recall and precision by averaging the percentage of

correctly retrieved (for recall) and correctly predicted (for precision) substituents per variable group, and then aggregate these averages across all variable groups. Finally, we report the Markush structure exact match, which requires both the CXSMILES exact match and substituent table exact match to be correct.

5.3. State-of-the-art Comparison

5.3.1. Multi-modal Markush Structure Recognition

Table 1 compares the MMSR methods for both visual and textual recognition of Markush structures across synthetic (MarkushGrapher-Synthetic) and real-world (M2S, USPTO-Markush) benchmarks. Our approach demonstrates superior performance to image-only, chemistry-specific models on most datasets. Notably, MarkushGrapher surpasses MolScribe in CXSMILES exact match accuracy on M2S (38% against 21%), USPTO-Markush (32% against 7%), and MarkushGrapher-Synthetic (65% against 11%). This substantial performance advantage highlights that DECIMER and MolScribe capture only a limited subset of Markush features, specifically R-groups represented as abbreviations. Interestingly, MolScribe achieves a higher CXSMILES Tanimoto score on the USPTO-Markush dataset. This score, which disregards Markush-specific structural features, highlights MolScribe’s robustness in identifying standard chemical structures. This observation inspired the integration of the OCSR encoder into MarkushGrapher.

Our method substantially outperforms general-purpose VLMs in CXSMILES recognition. VLMs fail to recognize Markush structures. Although Uni-SMART is trained to verify if a molecule is covered within a document containing Markush structures, it cannot actually identify these structures *per se*, a simpler task by comparison. MarkushGrapher also surpasses Uni-SMART in terms of the CXSMILES Tanimoto score. Furthermore, MarkushGrapher outperforms VLMs in table recognition, achieving higher exact match (29% vs. 19%) and F1-score (65% vs. 49%). It is worth noticing that the table exact match and F1-score

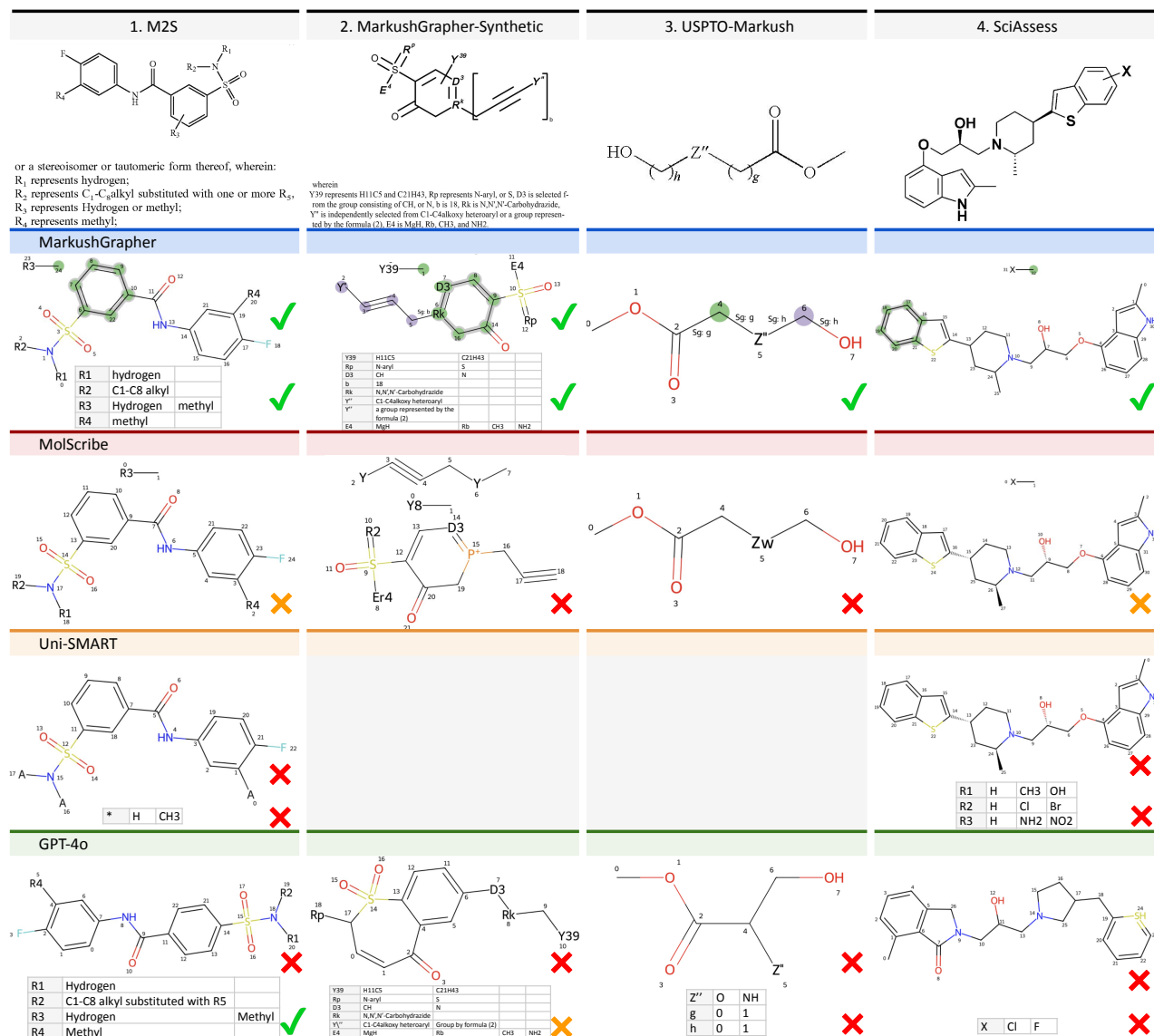


Figure 5. **Qualitative comparison.** Examples of predictions are shown for the different MMSR models on real-world data (M2S, USPTO-Markush, SciAssess) and on synthetic data (MarkushGrapher-Synthetic).

metrics may be disadvantageous for language models, as they often paraphrase content into semantically equivalent terms, which would be counted as errors; nonetheless, these metrics remain the most appropriate.

Overall, our multi-modal approach allows MarkushGrapher to outperform other methods on both the recognition of image and text in Markush structures.

5.4. Markush Features Analysis

Table 2 shows the performance of MarkushGrapher vs. image-only MMSR models for the different features of Markush structures. MarkushGrapher has a clearly higher CXSMILES exact match performance, largely because DECIMER and MolScribe can not generate predictions for

the ‘m’ and ‘Sg’ sections. While MarkushGrapher beats MolScribe at recognizing R-groups in the M2S dataset, it fails to do so in the USPTO-Markush dataset. This is probably due to the images of the M2S benchmark containing text descriptions, whereas USPTO-Markush consists solely of Markush backbone images. The performance of MarkushGrapher is notably stronger than the other models on USPTO-Markush, with its higher proportion of images with ‘Sg’ and ‘m’ sections. Additionally, MarkushGrapher is more effective at recognizing R-groups and ‘m’ sections compared to ‘Sg’ sections. Diversifying the frequency variation indicators in synthetic examples could further improve the recognition of ‘Sg’ sections.

Table 2. **Comparison of Markush features performances.** Impact of the Markush features on the recognition performances. The percentage of correctly recognized R-groups (R), ‘m’ sections (m), ‘Sg’ sections (Sg), and CXSMILES (EM) are reported.

Methods	M2S				USPTO-Markush			
	R	m	Sg	EM	R	m	Sg	EM
DECIMER [41]	12	0	0	3	34	0	0	7
MolScribe [38]	61	0	0	21	76	0	0	7
MarkushGrapher (Ours)	75	76	31	38	69	67	18	32

5.5. Qualitative Evaluation

Next, we conduct a qualitative evaluation of MarkushGrapher vs. SOTA methods. Figure 5 shows examples of predicted molecules for images from various benchmark datasets. MarkushGrapher accurately recognizes complex Markush image features such as position variation indicators (Figure 5, column 1, 2 and 4) and frequency variation indicators (Figure 5, column 2 and 3). Contrary to UniSMART and GPT-4o our model does not predict an incorrect table when the input image does not contain any table (Figure 5, column 3 and 4). Even when Markush features are disregarded in the evaluation, MarkushGrapher sometimes outperforms MolScribe (Figure 5, column 3), likely because MolScribe struggles with text descriptions or the presence of Markush features within the images. In addition, MarkushGrapher can handle the prediction of long tables which often confuse other models. (Cf. suppl. material for more details.)

5.6. Ablation Study

MarkushGrapher encoder. Table 3 shows the impact of the MarkushGrapher encoder. It can be observed that using the OCSR encoder, in addition, to the VTL encoder give an improvement of performances on the USPTO-Markush dataset. We also compare two fusion alternatives. Early fusion (EF) corresponds to concatenating the OCSR encodings with the input of the VTL encoder, and late fusion (LF)

Table 3. **MarkushGrapher encoder analysis.** Comparison of encoders in MarkushGrapher: VTL only, VTL and OCSR with early fusion (EF) or late fusion (LF).

Methods	M2S		USPTO-Markush	
	EM	T	EM	T
VTL	38	77	23	71
VTL + OCSR (EF)	38	77	28	76
VTL + OCSR (LF)	38	76	32	74

Table 4. **Markush structure representation analysis.** Performance of MarkushGrapher using optimized CXSMILES (Optimized), optimized CXSMILES without the R-groups compression (Compression), and optimized CXSMILES without the atom indexing (Indexing).

Methods	M2S		USPTO-Markush	
	EM	T	EM	T
Optimized	38	76	32	74
- Compression	30	68	31	70
- Indexing	35	71	24	79

corresponds to concatenating the OCSR encodings with the output of the VTL encoder. Late fusion yields better results. **Optimized Markush structure representation.** Table 4 highlights the impact of Markush structure representation. Removing the R-group compression or the atom indexing both lower performances on M2S and USPTO-Markush. The indexing is crucial for USPTO-Markush, which contains more CXSMILES with ‘m’ or ‘Sg’ sections.

Modalities dependencies. MarkushGrapher exploits dependencies between the text and image definitions of Markush structures. As shown in Figure 6, the model successfully infers a bond masked by a black patch in the input image. To deduce this bond, the model refers to the text description and applies chemistry rules. Specifically, if B” is an oxygen atom O, there must be two connections to the rest of the molecule, indicating a double bond. Conversely, if B” is an alcohol group OH, there only must be a single connection, resulting in a single bond. This ability to understand cross-modal dependencies is enabled by our joint visual and textual recognition approach, which is helpful for accurately recognizing complex multi-modal Markush structures. (Cf. suppl. material for further analysis.)

6. Conclusion

We propose a novel architecture to recognize multi-modal Markush structures in documents by combining a Vision-Text-Layout encoder and an Optical Chemical Structure Recognition encoder. Our model jointly process visual and textual definitions of a Markush structure and converts these to a structured graph and table. The model is trained on synthetic images and demonstrates strong generalization capabilities, allowing it to outperform existing methods on real-world data including our new M2S dataset. MarkushGrapher is a step towards large-scale extraction of Markush structures in documents, a key challenge in patent analysis [42].

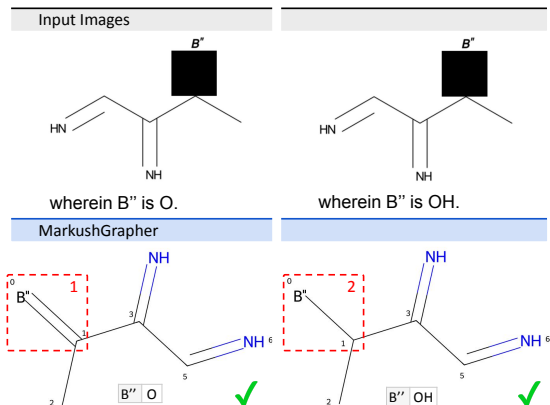


Figure 6. **Cross-modality understanding.** A bond is masked in the input image (black patch). For an oxygen atom O, MarkushGrapher predicts a double bond (red box 1). For an alcohol group OH, MarkushGrapher predicts a single bond (red box 2).

References

- [1] ChemAxon Marvin JS. <https://marvinjs-demo.chemaxon.com/latest/demo.html>. Accessed: January 2025. 5
- [2] United States Patent and Trademark Office. <http://uspto.gov>. Accessed: January 2025. 6
- [3] ChemAxon Extended SMILES (CXSMILES) Documentation. https://docs.chemaxon.com/display/docs/formats_chemaxon-extended-smiles-and-smarts-cxsmiles-and-cxsmarts. Accessed: January 2025. 4
- [4] Praveesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, et al. Pixtral 12B. *arXiv preprint arXiv:2410.07073*, 2024. 6
- [5] Christoph Auer, Maksym Lysak, Ahmed Nassar, Michele Dolfi, Nikolaos Livathinos, Panos Vagenas, Cesar Berrospi Ramis, Matteo Omenetti, Fabian Lindlbauer, Kasper Dinkla, et al. Docling Technical Report. *arXiv preprint arXiv:2408.09869*, 2024. 3
- [6] John M. Barnard, Michael F. Lynch, and Stephen M. Welford. Computer storage and retrieval of generic chemical structures in patents. 2. GENSAL, a formal language for the description of generic chemical structures. *Journal of Chemical Information and Computer Sciences*, 21(3):151–161, 1981. 5
- [7] 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(4): 2059–2072, 2020. 3
- [8] Hengxing Cai, Xiaochen Cai, Junhan Chang, Sihang Li, Lin Yao, Changxin Wang, Zhifeng Gao, Hongshuai Wang, Yongge Li, Mujie Lin, et al. SciAssess: Benchmarking LLM Proficiency in Scientific Literature Analysis. *arXiv preprint arXiv:2403.01976*, 2024. 5, 6
- [9] Hengxing Cai, Xiaochen Cai, Shuwen Yang, Jiankun Wang, Lin Yao, Zhifeng Gao, Junhan Chang, Sihang Li, Mingjun Xu, Changxin Wang, et al. Uni-SMART: Universal Science Multimodal Analysis and Research Transformer. *arXiv preprint arXiv:2403.10301*, 2024. 2, 3, 6
- [10] Yufan Chen, Ching Ting Leung, Yong Huang, Jianwei Sun, Hao Chen, and Hanyu Gao. MolNexTR: A Generalized Deep Learning Model for Molecular Image Recognition. *arXiv preprint arXiv:2403.03691*, 2024. 2
- [11] Arthur Dalby, James G. Nourse, W. Douglas Hounshell, Ann K. I. Gushurst, David L. Grier, Burton A. Leland, and John Laufer. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *Journal of Chemical Information and Computer Sciences*, 32(3):244–255, 1992. 5
- [12] Winfried Dethlefsen, Michael F. Lynch, Valerie J. Gillet, Geoffrey M. Downs, John D. Holliday, and John M. Barnard. Computer storage and retrieval of generic chemical structures in patents. 11. Theoretical aspects of the use of structure languages in a retrieval system. *Journal of Chemical Information and Computer Sciences*, 31(2):233–253, 1991. 1
- [13] Tommy. Ebe, Karen A. Sanderson, and Patricia S. Wilson. The Chemical Abstracts Service generic chemical (Markush) structure storage and retrieval capability. 2. The MARPAT file. *Journal of Chemical Information and Computer Sciences*, 31(1):31–36, 1991. 2
- [14] 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. 2, 3
- [15] William Fisanick. The Chemical Abstract’s Service generic chemical (Markush) structure storage and retrieval capability. 1. Basic concepts. *Journal of Chemical Information and Computer Sciences*, 30(2):145–154, 1990. 2
- [16] Carina S. Haupt. Markush structure reconstruction. Master thesis, 2009. 3
- [17] Stephen R. Heller, Alan McNaught, Igor Pletnev, Stephen Stein, and Dmitrii Tchekhovskoi. InChI, the IUPAC International Chemical Identifier. *Journal of Cheminformatics*, 7(1):23, 2015. 6
- [18] Rodrigo Hormazabal, Changyoung Park, Soonyoung Lee, Sehui Han, Yeonsik Jo, Jaewan Lee, Ahra Jo, Seung Hwan Kim, Jaegul Choo, Moontae Lee, and Honglak Lee. CEDE: A collection of expert-curated datasets with atom-level entity annotations for Optical Chemical Structure Recognition. In *Advances in Neural Information Processing Systems*, pages 27114–27126. Curran Associates, Inc., 2022. 3
- [19] Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. mPLUG-DocOwl 1.5: Unified Structure Learning for OCR-free Document Understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3096–3120, Miami, Florida, USA, 2024. Association for Computational Linguistics. 3
- [20] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking. In *Proceedings of the 30th ACM International Conference on Multimedia*, page 4083–4091, New York, NY, USA, 2022. Association for Computing Machinery. 2, 3
- [21] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of Experts. *arXiv preprint arXiv:2401.04088*, 2024. 5
- [22] Ivan Khokhlov, Lev Krasnov, Maxim V. Fedorov, and Sergey Sosnin. Image2SMILES: Transformer-Based Molecular Optical Recognition Engine. *Chemistry-Methods*, 2(1): e202100069, 2022. 2
- [23] Geewook Kim, Teakgyu Hong, Moonbin Yim, Jeongyeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pages 498–517. Springer, 2022. 3

- [24] 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 2019 update: improved access to chemical data. *Nucleic Acids Research*, 47(D1):D1102–D1109, 2018. 4
- [25] Nikolay A. Kolpakov, Alexey I. Molodchenkov, and Anton V. Lukin. Methods of extracting biomedical information from patents and scientific publications (on the example of chemical compounds). *Discrete and Continuous Models and Applied Computational Science*, 31(1):64–74, 2023. 3
- [26] Iliia Korvigo, Maxim Holmatov, Anatolii Zaikovskii, and Mikhail Skoblov. Putting hands to rest: efficient deep CNN-RNN architecture for chemical named entity recognition with no hand-crafted rules. *Journal of Cheminformatics*, 10(1):28, 2018. 1
- [27] Greg Landrum. RDKit: Open-Source Cheminformatics Software. <http://www.rdkit.org/>. Accessed: January 2025. 4
- [28] Fan Lin and Jianhua Li. MPOCSR: optical chemical structure recognition based on multi-path Vision Transformer. *Complex & Intelligent Systems*, 10(6):7553–7563, 2024. 2
- [29] Daniel M. Lowe and Roger A. Sayle. LeadMine: a grammar and dictionary driven approach to entity recognition. *Journal of Cheminformatics*, 7(1):S5, 2015. 3
- [30] Lokesh Mishra, Cesar Berrospi, Kasper Dinkla, Diego Antognini, Francesco Fusco, Benedikt Bothur, Maksym Lysak, Nikolaos Livathinos, Ahmed Nassar, Panagiotis Vagenas, Lucas Morin, Christoph Auer, Michele Dolfi, and Peter Staar. ESG Accountability Made Easy: DocQA at Your Service. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(21):23814–23816, 2024. 3
- [31] 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, 2023. 1, 3
- [32] 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(1):6532, 2024. 1
- [33] Ahmed Nassar, Andres Marafioti, Matteo Omenetti, Maksym Lysak, Nikolaos Livathinos, Christoph Auer, Lucas Morin, Rafael Teixeira de Lima, Yusik Kim, A. Said Gurbuz, Michele Dolfi, Miquel Farré, and Peter W. J. Staar. SmolDocling: An ultra-compact vision-language model for end-to-end multi-modal document conversion. *arXiv preprint arXiv:2503.11576*, 2025. 3
- [34] Martijn Oldenhof, Adam Arany, Yves Moreau, and Edward De Brouwer. Weakly Supervised Knowledge Transfer with Probabilistic Logical Reasoning for Object Detection. In *The Eleventh International Conference on Learning Representations*, 2023. 2
- [35] Martijn Oldenhof, Edward De Brouwer, Adam Arany, and Yves Moreau. Atom-Level Optical Chemical Structure Recognition with Limited Supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17669–17678, 2024. 2
- [36] OpenAI. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*, 2024. 6
- [37] Edward O. Pyzer-Knapp, Matteo Manica, Peter Staar, Lucas Morin, Patrick Ruch, Teodoro Laino, John R. Smith, and Alessandro Curioni. Foundation models for materials discovery – current state and future directions. *npj Computational Materials*, 11(1):61, 2025. 1
- [38] 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. 1, 2, 3, 5, 6, 8
- [39] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1), 2020. 5
- [40] Kohulan Rajan, Henning Otto Brinkhaus, Maria Sorokina, Achim Zielesny, and Christoph Steinbeck. DECIMER-Segmentation: Automated extraction of chemical structure depictions from scientific literature. *Journal of Cheminformatics*, 13(1):20, 2021. 1
- [41] Kohulan Rajan, Henning Otto Brinkhaus, M. Isabel Agea, Achim Zielesny, and Christoph Steinbeck. DECIMER.ai: an open platform for automated optical chemical structure identification, segmentation and recognition in scientific publications. *Nature Communications*, 14(1):5045, 2023. 1, 2, 6, 8
- [42] Edlyn S. Simmons. Markush structure searching over the years. *World Patent Information*, 25(3):195–202, 2003. 1, 8
- [43] 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. 1
- [44] Daylight Chemical Information Systems. Daylight Theory Manual. 2011. 6
- [45] Bowen Tang, Zhangming Niu, Xiaofeng Wang, Junjie Huang, Chao Ma, Jing Peng, Yinghui Jiang, Ruiquan Ge, Hongyu Hu, Luhao Lin, and Guang Yang. Automated molecular structure segmentation from documents using ChemSAM. *Journal of Cheminformatics*, 16(1):29, 2024. 1
- [46] Zineng Tang, Ziyi Yang, Guoxin Wang, Yuwei Fang, Yang Liu, Chenguang Zhu, Michael Zeng, Cha Zhang, and Mohit Bansal. Unifying Vision, Text, and Layout for Universal Document Processing. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 19254–19264. IEEE, 2023. 2, 3, 4
- [47] T.T. Tanimoto. *An Elementary Mathematical Theory of Classification and Prediction*. International Business Machines Corporation, 1958. 6
- [48] Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. Label Studio: Data labeling software. <https://github.com/heartexlabs/label-studio>. Accessed: January 2025. 5
- [49] Aniko T. Valko and A. Peter Johnson. CLiDE Pro: The Latest Generation of CLiDE, a Tool for Optical Chemical

- Structure Recognition. *Journal of Chemical Information and Modeling*, 49(4):780–787, 2009. [2](#), [3](#)
- [50] Stefan Wagner, Christian Sternitzke, and Sascha Walter. Mapping Markush. *Research Policy*, 51(10):104597, 2022. [3](#)
- [51] Jie Wang, Zihao Shen, Yichen Liao, Zhen Yuan, Shiliang Li, Gaoqi He, Man Lan, Xuhong Qian, Kai Zhang, and Honglin Li. Multi-modal chemical information reconstruction from images and texts for exploring the near-drug space. *Briefings in Bioinformatics*, 23(6):bbac461, 2022. [3](#), [4](#)
- [52] Haoran Wei, Chenglong Liu, Jinyue Chen, Jia Wang, Lingyu Kong, Yanming Xu, Zheng Ge, Liang Zhao, Jianjian Sun, Yuang Peng, et al. General OCR Theory: Towards OCR-2.0 via a Unified End-to-end Model. *arXiv preprint arXiv:2409.01704*, 2024. [2](#), [3](#)
- [53] David Weininger. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 1988. [2](#)
- [54] Egon L. Willighagen, John W. Mayfield, Jonathan Alvarsson, Arvid Berg, Lars Carlsson, Nina Jeliakova, Stefan Kuhn, Tomáš Pluskal, Miquel Rojas-Chertó, Ola Spjuth, et al. The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. *Journal of Cheminformatics*, 9(1):33, 2017. [4](#)
- [55] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, et al. UReader: Universal OCR-free Visually-situated Language Understanding with Multimodal Large Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2841–2858, Singapore, 2023. Association for Computational Linguistics. [3](#)
- [56] Zenan Zhai, Dat Quoc Nguyen, Saber Akhondi, Camilo Thorne, Christian Druckenbrodt, Trevor Cohn, Michelle Gregory, and Karin Verspoor. Improving Chemical Named Entity Recognition in Patents with Contextualized Word Embeddings. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 328–338, Florence, Italy, 2019. Association for Computational Linguistics. [1](#)
- [57] Chong Zhou, Wei Liu, Xiyue Song, Mengling Yang, and Xiaowang Peng. YoDe-Segmentation: automated noise-free retrieval of molecular structures from scientific publications. *Journal of Cheminformatics*, 15(1):111, 2023. [1](#)