MarkushGrapher-2: End-to-end Multimodal Recognition of Chemical Structures

Tim Strohmeyer [1, 2] Lucas Morin[1, 2] Gerhard Ingmar Meijer[1] Valéry Weber[1] Ahmed Nassar[1]
Peter Staar[1]
[1]IBM Research [2]ETH Zurich
{tis, lum, inm, vwe, ahn, taa}@zurich.ibm.com

Abstract

Automatically extracting chemical structures from documents is essential for the large-scale analysis of the literature in chemistry. Automatic pipelines have been developed to recognize molecules represented either in figures or in text independently. However, methods for recognizing chemical structures from multimodal descriptions (Markush structures) lag behind in precision and cannot be used for automatic large-scale processing. In this work, we present MarkushGrapher-2, an end-to-end approach for the multimodal recognition of chemical structures in documents. First, our method employs a dedicated OCR model to extract text from chemical images. Second, the text, image, and layout information are jointly encoded through a Vision-Text-Layout encoder and an Optical Chemical Structure Recognition vision encoder. Finally, the resulting encodings are effectively fused through a two-stage training strategy and used to auto-regressively generate a representation of the Markush structure. To address the lack of training data, we introduce an automatic pipeline for constructing a large-scale dataset of real-world Markush structures. In addition, we present IP5-M, a large manually-annotated benchmark of real-world Markush structures, designed to advance research on this challenging task. Extensive experiments show that our approach substantially outperforms state-of-the-art models in multimodal Markush structure recognition, while maintaining strong performance in molecule structure recognition. Code, models, and datasets are released publicly.[1]

1. Introduction

Extracting chemical structures from documents is essential for unifying knowledge in chemistry. It enables search engines to retrieve information using molecular queries and allows the creation of training datasets for machine learning models. Converting unstructured documents into machine-readable formats can ultimately accelerate discovery across the life and materials sciences [20]. Several automated pipelines have been developed to create molecular databases from unstructured documents [17, 23, 24]. These systems extract molecules either from their textual definitions, using chemical named entity recognition [15, 28, 34], or from their visual representations, using chemical structure image segmentation [22, 29, 35] and recognition [16, 21, 23]. More recently, progress in multimodal document understanding has enabled extracting more complex chemical representations that combine visual and textual information, known as Markush structures [18]. These widely used representations provide compact descriptions of families of related molecules. A Markush structure includes both an image and a text component: the image defines the

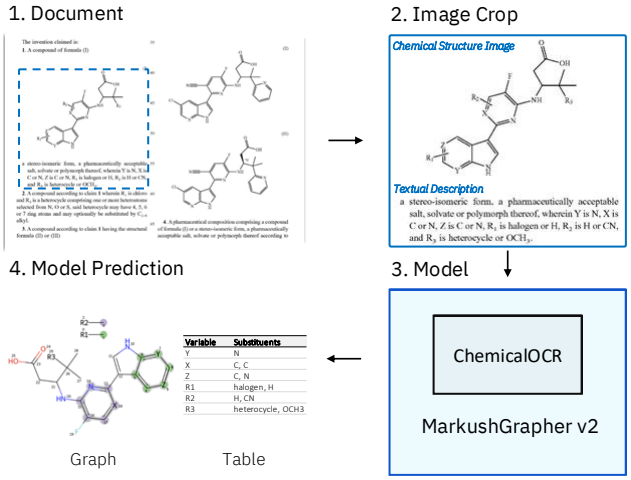

Figure 1. **Model Use Case**: MarkushGrapher-2 parses Markush backbones and variable regions from document image crops via joint multimodal encoding of vision, text, and layout.

[1]https://github.com/DS4SD/MarkushGrapher

Markush backbone, containing atoms, bonds, and variable regions, while the text specifies the molecular substituents that can replace those variable regions. The variable regions may include variable groups (also named residual R-groups), frequency variation indicators, and positional variation indicators [5]. Markush structures play a central role in patent analysis, supporting prior-art searches, freedom-to-operate evaluations, and landscape analysis [26]. Despite their importance in chemical research, their coverage in databases remains limited. Currently, Markush structures are only indexed in the proprietary manually-created databases MARPAT [6] and DWPIM [11].

Multi-modal Markush Structure Recognition (MMSR) poses key challenges that limit the performance of their automatic recognition. First, images of Markush structure backbones follow a wide range of conventions and drawing standards. In patent documents, for example, the visual style can vary substantially across patent offices and publication years. Second, the textual definitions of Markush structures lack standardization and often contain condition-based or recursive descriptions. Third, there is a lack of real-world training datasets with comprehensive annotations of Markush structure visual and textual definitions.

In this work, we introduce MarkushGrapher-2, a model for end-to-end multimodal recognition of Markush structures, illustrated in Figure 1. MarkushGrapher-2 extends the MarkushGrapher [18] framework into a universal approach capable of recognizing both molecular images and multimodal Markush structures. The model follows an encoder–decoder architecture that takes as input an image of a molecule or Markush structure and outputs a textual sequence representing its structure. This output is divided into two components: (1) a graph of the Markush backbone, and (2) a table of possible substituents that replace the variable groups in the backbone. The input image is jointly encoded using two encoders, a Vision-Text-Layout (VTL) encoder and a vision encoder, pretrained for the task of Optical Chemical Structure Recognition (OCSR). These encodings are projected, concatenated and fed to a text decoder to autoregressively generate a sequential Markush representation. Compared to its predecessor, MarkushGrapher-2 introduces several major improvements. First, it integrates a dedicated OCR module, enabling end-to-end processing. Here, *end-to-end* refers to the model's ability to directly process a raw image at inference time, without requiring pre-annotated OCR outputs. Second, it adopts a new two-phase training strategy designed to improve the fusion of encoders. Third, the model is trained using a new training data generation pipeline, allowing it to recognize both molecule images and multi-modal Markush structures. To further support research in this area, we introduce IP5-M, a benchmark dataset of manually annotated Markush structures extracted from patent documents of the IP5 patent offices, United States Patent and Trademark Office (USPTO), Japan Patent Office (JPO), Korean Intellectual Property Office (KIPO), China National Intellectual Property Administration (CNIPA), and European Patent Office (EPO). Comprehensive experiments across multiple benchmarks for MMSR demonstrate that MarkushGrapher-2 consistently outperforms both general-purpose and chemistry-specific document understanding models, while maintaining competitive performance on OCSR benchmarks. In summary:

- We develop MarkushGrapher-2, a universal model for recognizing both molecular images and multi-modal Markush structures.
- We introduce a dedicated ChemicalOCR module for end-to-end processing and improved abbreviation recognition.
- We design a two-phase training strategy to improve MarkushGrapher-2's encoder fusion.
- We create a data generation pipeline for real-world Markush backbone training samples from MOL files and accompanying images provided by the USPTO.
- We release IP5-M, a manually annotated benchmark of real-world Markush structures from IP5 patent offices.

2. Related Work

Most existing approaches to Markush structure recognition focus exclusively on the image component, neglecting supporting information through textual descriptions. Some methods, originally designed for Optical Chemical Structure Recognition (OCSR) methods, can identify a limited subset of Markush structure backbones, i.e., chemical structures containing variable groups [3, 21, 23]. More recently, MolParser extended this capability to capture positional and frequency variation indicators [8]. However, its handling of frequency variation is restricted to single-atom cases only. Overall, these approaches remain limited in both scope and accuracy. Multimodal pipelines for chemical document understanding have shown progress [7, 31]. General-purpose vision–language models have also begun to incorporate molecule-image recognition as one of their tasks, such as GOT-OCR 2.0 [32], Qwen2.5-VL [1], PaliGemma 2 [27], and DeepSeek-OCR [33]. Their prediction is however limited. Domain-specific VLMs like Uni-SMART [2] and PatentFinder [25] can also determine whether a query molecule is covered by a Markush definition in a document, which implicitly requires multimodal Markush recognition. However, these models are not practical for searching large corpuses, as each query would require running the model across all candidate documents. MarkushGrapher [18] is the only vision-language model capable of jointly interpreting textual and visual components of Markush definitions. Its initial version requires access to all OCR cells in the input image, which is a limitation for an integration in an end-to-end processing pipeline; and its visual recognition accuracy has room for improvement.

3. MarkushGrapher-2

MarkushGrapher-2 is a transformer-based model that jointly encodes vision, text, and layout modalities for multimodal chemical structure recognition. The model integrates two complementary encoders: a Vision Encoder, pretrained for standard Optical Chemical Structure Recognition (OCSR), and a Vision–Text–Layout (VTL) Encoder, trained specifically for Markush features extraction. The latter leverages textual and positional information present in chemical documents—such as variable groups, positional and frequency variation indicators—to produce a unified multimodal representation.

3.1. Architecture

Figure 2 illustrates the model architecture, which consists of two complementary encoding pipelines. In the first pipeline, the input image is processed by a vision encoder (taken from MolScribe [21]) pretrained for OCSR, adopting a Swin-B ViT backbone to extract visual features representing molecular structures. In the second pipeline, the same image is passed through an OCR module that detects and recognizes textual elements within the image. These elements include atom labels, abbreviations, and descriptive text (e.g., variable group definitions) typically located near or below the structure. The extracted text and bounding boxes are combined with the image patches and fed into a VTL encoder based on a T5-base backbone. Following the UDOP fusion paradigm [30], visual and textual tokens that spatially coincide—according to their bounding boxes—are aligned and fused to form a combined multimodal representation. The output of the VTL encoder is then concatenated with a projected embedding from the pretrained Vision Encoder. This joint representation is passed to a text decoder, which autoregressively generates a structured sequence describing the chemical backbone and its associated Markush features. The generated Markush features include variable groups, frequency variation indicators, positional variation indicators, and a table containing the variable group substituents mentioned in the textual description.

3.2. OCR Model

The Optical Character Recognition (OCR) module is a critical component of the model architecture, providing the textual and layout modalities necessary for interpreting complex Markush features. The OCR module extracts character-level text and bounding boxes from the input image. Subsequently, the extracted text, bounding boxes and vision patches are passed to the model's VLT encoder for multimodal encoding.

Due to limited performance of existing OCR models on chemical images, we introduce ChemicalOCR, a compact vision–language model (VLM) fine-tuned for OCR in chemical images. The ChemicalOCR architecture is based on Smoldocling [19], a lightweight 256M-parameter model originally developed for end-to-end document conversion.

3.3. Two-stage Training Strategy

To train MarkushGrapher-2, we adopt a two-stage training strategy designed to fully leverage the pretrained features of the vision encoder for OCSR, while effectively fusing them with the multimodal representations learned by the VTL encoder for MMSR. Figure 3 shows a depiction of the two training phases.

During Phase 1 (Adaptation), the Vision Encoder, Projector, and Text Decoder are trained in isolation for the task of standard SMILES (Simplified Molecular Input Line Entry System) prediction. Keeping the Vision Encoder weights frozen, this allows the decoder and projector to adapt to the pretrained OCSR feature space without altering the original visual representations. During Phase 2 (Fusion), the OCR model and VTL encoder are introduced to the model architecture. The Vision Encoder, Projector, and OCR model are frozen and the VTL encoder and Text decoder are trained end-to-end for the task of CXSMILES (Chemaxon Extended SMILES) and Substituent Table prediction.

This two-phase training strategy allows the model to effectively leverage the pretrained Vision Encoder and build upon the OCSR features to perform the more challenging task of Markush structure recognition. By freezing the Vision Encoder and Projector during Phase 2 we preserve the original OCSR feature space and ensure that the VTL encoder focuses on learning the missing features required for Markush structure recognition (i.e., CXSMILES and substitutent table prediction).

4. Datasets

MarkushGrapher-2 is trained on diverse sets of synthetic data and real-world data. The data is sourced and converted from different public document collections and databases: IP5 patent offices, PubChem, and datasets published by MolScribe [21] and MolParser [9].

4.1. The USPTO-MOL-M Dataset

For the training of MarkushGrapher-2, we introduce a new large-scale training dataset, consisting of real-world image-(CX)SMILES pairs extracted automatically from MOL files that are provided by the USPTO along with the patent document. The USPTO MOL files (V2000 format) contain information such as atom symbols, atomic coordinates, bonds and bond types, aliases (abbreviations and R-groups), superatoms (abbreviations), and frequency variation indicators. These MOL files do not include all the visual details that are shown in their corresponding image, for example, positional variation indicators are not directly included. We developed code to clean the MOL files and convert them

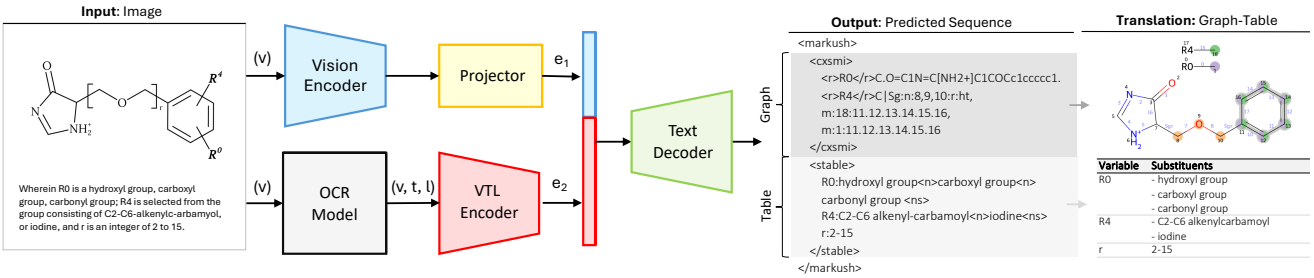

Figure 2. **Model Architecture**: MarkushGraher-2 employs two complementary encoding pipelines. In the first pipeline, the input image is processed by a vision encoder (blue) followed by an MLP projector (yellow). In the second pipeline, the image is passed through an OCR model to extract textual content and bounding boxes, which is then fed into a Vision–Text–Layout (VTL) encoder together with the original image. The output of the MLP projector (e1) is concatenated with the resulting VTL embedding (e2). The combined representation is passed to a text decoder to generate a sequential description of the Markush structure and its substituents in tabular form.

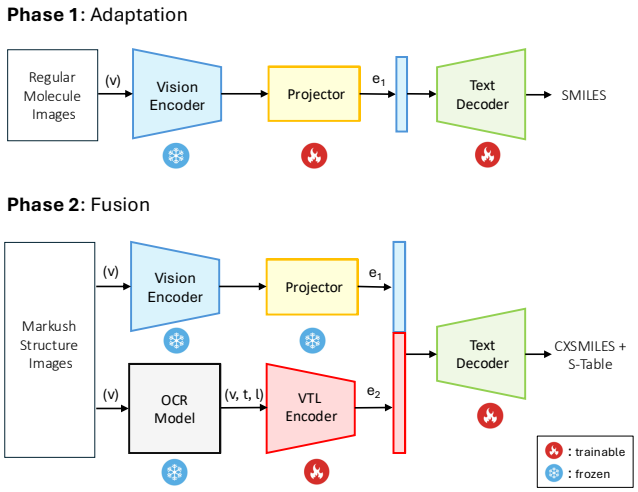

Figure 3. **Two-Phase Training**: In Phase 1 (Adaptation), the OCSR encoder is frozen while the MLP projector and text decoder are trained for SMILES prediction to align with pretrained OCSR features. In Phase 2 (Fusion), the adapted modules are initialized, the VTL encoder is introduced, and the full model is trained end-to-end for CXSMILES prediction.

to CXSMILES format. In this process, positional variation indicators were reconstructed based on atomic coordinates and bonding patterns, frequency variation indicators were extracted from the structural repeating unit sections, and R-groups were derived from the alias and superatom sections. Unnecessary elements—such as text, labels, or indices present in the MOL files—were removed to the greatest extent possible. The dataset built in this study comprises USPTO MOL files from the years 2010 to 2025.

4.2. Training Datasets

The datasets used for training MarkushGrapher-2 can be categorized into three categories: (1) OCR data, consisting of image-OCR pairs for optical character recognition on chemical structure images, (2) OCSR data, consisting of image–SMILES pairs for standard molecular structure recognition, and (3) MMSR data, comprising image–CXSMILES pairs (and substituent tables) for recognizing Markush structures.

ChemicalOCR module is pretrained on a set of 235k synthetic chemical structures containing automatic OCR annotations. To generate the training images, SMILES are randomly sampled from PubChem [14] and augmented into chemically valid CXSMILES representations, introducing variable groups, frequency and positional variation indicators. These CXSMILES are then rendered into images containing corresponding molecular drawings, textual annotations, and character-level bounding boxes [18]. To substantially improve model performance on real-world chemical images, the model is further finetuned using a set of 7k manually annotated OCR samples with chemical structures cropped from IP5 patent documents.

During Phase 1 (Adaptation), the model is trained exclusively on OCSR data. Specifically, we use 243k real-world image–SMILES pairs sourced from the public MolScribe dataset [21], derived from USPTO documents.

In the subsequent Phase 2 (Fusion), we train the model on a combination of MMSR datasets. This includes same 235k synthetically generated image–CXSMILES pairs with corresponding substituent tables that were used for training the ChemicalOCR module. Additionally, the MMSR training corpus includes 91k real-world Markush samples from the MolParser dataset [9], that are converted into an optimized CXSMILES prediction format, and 54k real-world Markush samples from the USPTO-MOL-M dataset described above.

5. Experiments

5.1. Implementation Details

The Vision-Text-Layout encoder and text decoder are based on a T5 encoder–decoder architecture [30]. The OCSR en-

coder is the vision encoder taken from MolScribe, which is based on a SWIN transformer architecture [21]. This encoder remains frozen during training. In total, the model comprises 831M parameters, of which 744M are trainable. Training is performed in two phases. During Phase 1 (Adaptation), the Projector and Text Decoder are trained on 243K real samples for 3 epochs using an NVIDIA A100 GPU. Here, we use the Adam optimizer with a learning rate of 5e-4, 1000 warm-up steps, a batch size of 10, and a weight decay of 1e-3. For Phase 2 (Fusion), the pretrained Vision encoder and Projector are frozen, and the VTL encoder and text decoder are further trained using a mix of 235k synthetic and 145k real-world Markush structure samples. Here we use a batch size of 8 and train for 2 epochs.

5.2. Evaluation Datasets and Metrics

We evaluate MarkushGrapher-2 on two tasks: (1) image recognition and (2) substituent table recognition.

Markush Benchmarks: To evaluate model performance for Markush structure recognition under diverse real-world conditions, we employ several benchmark datasets of Markush structures. M2S contains 103 real-world Markush structure images with textual descriptions, manually annotated and substituent tables [18]. USPTO-M contains 74 real-world Markush structure images, manually annotated [18]. WildMol-M, introduced by MolParser [9], contains 10k real-world, semi-manually annotated Markush structure images. Finally, we introduce a new benchmark, IP5-M, consisting of 1,000 manually annotated Markush structures from patent documents of the IP5 offices published between 1980 and 2025.

OCSR Benchmarks: To evaluate model performance on standard SMILES prediction, we use four common OCSR benchmarks, USPTO [10], JPO [12], UOB [12], and WildMol [9].

OCR Benchmarks: The above M2S, USPTO-M, and IP5-M benchmarks are also used to evaluate the ChemicalOCR performance on OCR predictions.

CXSMILES Accuracy (A): Image recognition performance is measured using the CXSMILES accuracy (A), while substituent table recognition is evaluated using prediction accuracy and F1 score. Markush accuracy measures the overall accuracy of correct Table and CXSMILES predictions. Stereochemistry is ignored during evaluation to ensure consistency across datasets. In more detail, the CXSMILES accuracy (A) metric quantifies the percentage of perfectly recognized molecular representations. A prediction is considered correct if two conditions are satisfied: (1) the predicted SMILES (without Markush features) corresponds to the ground truth according to `InChIKey` equivalence, and (2) the Markush features, i.e., variable groups (R-groups), positional and frequency variation indicators are correctly represented.

5.3. Evaluation and State-of-the-art Comparison

In this section we provide an evaluation of the ChemicalOCR module and the overall MarkushGrapher-2 performance versus state-of-the-art methods.

Optical Character Recognition: Table 1 shows a quantitative comparison of our ChemicalOCR model with PaddleOCR v5 [4] and EasyOCR [13]. ChemicalOCR substantially outperforms PaddleOCR and EasyOCR. Figure 4 presents a visual comparison of OCR predictions obtained by each evaluated method for a representative Markush structure image from each benchmark. Consistent with the quantitative results in Table 1, ChemicalOCR demonstrates superior performance in both character localization and recognition. It accurately identifies long abbreviations within chemical structures and correctly parses textual descriptions below the images (see M2S, second row). It is observed that EasyOCR struggles with longer text sequences, including abbreviations and descriptive text. PaddleOCR v5 handles longer text sequences, such as descriptive captions, reliably, but often misinterprets characters within chemical structures—frequently merging symbols into a single bounding box and confusing bonds with minus or equal signs.

Markush Structure Recognition: Table 2 shows a quantitative comparison of our MarkushGrapher-2 model with image-only models MolParser-Base and MolScribe, and multimodel models GPT-5, DeepSeek-OCR, and MarkushGrapher-1. GPT-5 was run on the M2S benchmark, which is reasonably representative for Markush structure recognition. MarkushGrapher-2 substantially outperforms state-of-the-art models on Markush structure recognition. Figure 5 presents a visual comparison of predictions of MarkushGrapher-2 and state-of-the-art models for a representative Markush structure image from each benchmarks. As illustrated, MarkushGrapher-2 accurately reconstructs both the molecular backbone from the structural image, as well as complex Markush-specific features, such as variable groups, positional and frequency variation indicators. Specifically, the model effectively captures any-locant cycle connections, repeating structural units, and end-to-molecule attach points (AP). Note: Scores for MarkushGrapher-1 cannot be provided for WildMold-M, since MarkushGrapher-1 does not constitute an OCR module.

Molecular Structure Recognition: Table 3 shows a quantitative comparison of our MarkushGrapher-2 model with image-only models MolParser-Base and MolScribe, and multimodel models GPT-5, DeepSeek-OCR, and MarkushGrapher-1. GPT-5 was run on the JPO benchmark, which is reasonably representative for molecular structure recognition. The results show that MarkushGrapher-2 is competitive with state-of the-art models on molecular structure recognition.

Models	M2S (103)				USPTO-M (74)				IP5-M (1000)			
	P	R	F1	A@IoU_0.5	P	R	F1	A@IoU_0.5	P	R	F1	A@IoU_0.5
PaddleOCR v5	8.9	6.8	7.7	0.0	2.3	1.1	1.2	0.0	2.2	1.7	1.9	0.6
EasyOCR	9.8	10.7	10.2	0.0	24.8	14.2	18.0	0.0	23.5	15.2	18.4	2.7
ChemicalOCR (Ours)	**86.9**	**87.4**	**87.2**	**32.0**	**93.5**	**92.6**	**93.0**	**63.5**	**85.6**	**87.4**	**86.5**	**69.5**

Table 1. **OCR on Chemical Images**: Comparison of our ChemicalOCR model with existing OCR models. The evaluation is conducted on real-world benchmarks (M2S, USPTO-M, and IP5-M). Precision P, Recall R, and F1 are measured at individual bounding-box level. Accuracy A is measured at the image level; an image is considered correct if all OCR cells have an IoU > 0.5 and their recognized characters match the ground truth.

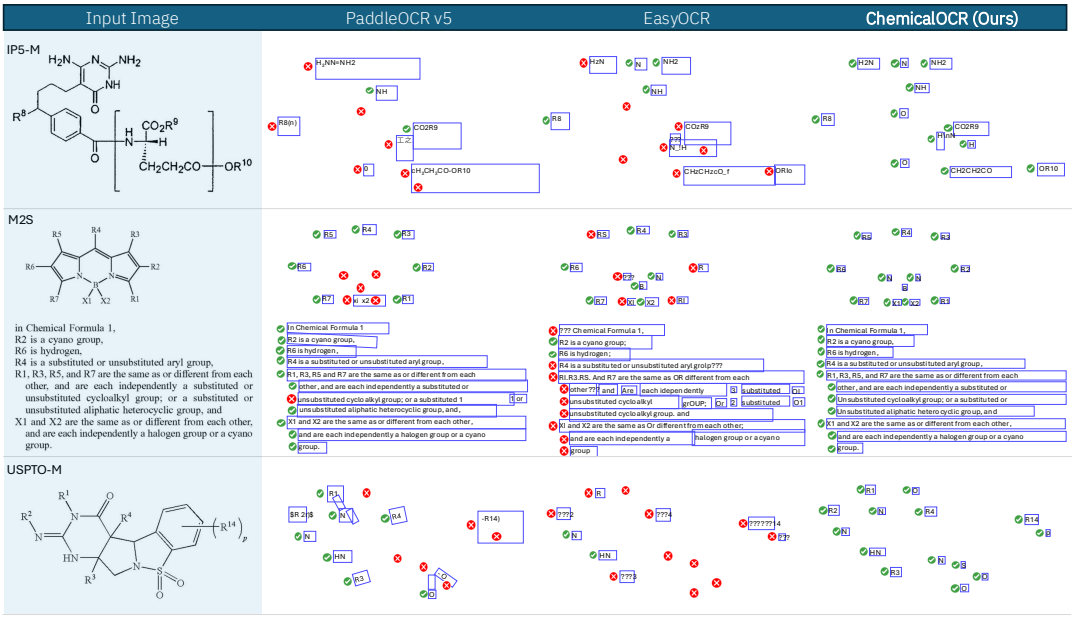

Figure 4. **OCR - Qualitative Comparison**: Comparison of OCR predictions by three models PaddleOCR v5, EasyOCR, and ChemicalOCR (Ours) for an exemplary chemical structure from the benchmarks M2S, USPTO-M, and IP5-M. Red labels indicate incorrect OCR, green labels indicate correct OCR and blue indicates predicted bounding boxes.

5.4. Ablation Study

Effect of OCR module: Table 4 compares the performance of MarkushGrapher-2 with and without OCR predictions from the ChemicalOCR module as input. The scores provide insight about the importance of the text and layout modality for the overall model prediction. We observe that the OCR predictions substantially improves MarkushGrapher-2 prediction accuracy. Figure 6 illustrates an example prediction from MarkushGrapher-2 with and without OCR input. While MarkushGrapher-2 without OCR input accurately predicts the structural backbone, it fails to capture the Markush features, i.e., the repeating groups (Sg). It may be inferred that the text extracted by the OCR—such as brackets and indices—provides important additional information that substantially improves Markush feature prediction.

Effect of Two-Phase Training: Table 5 shows the impact of the proposed two-phase training on overall MarkushGrapher-2 performance. We compare the two-phase setup—first adapting the Projector and VTL decoder to the pretrained OCSR features, followed by joint training of both encoders—with a single-phase training approach. In the table, the single-phase and two-phase setups are labeled with 'Fusion only' and 'Adaptation and Fusion', respectively. This experiment highlights the benefits of adapting the decoder to the pretrained OCSR features before fusing encoder outputs for Markush structure recognition. The results show that two-phase training improves the model's ability to encode Markush features while preserving performance on standard molecular recognition.

Methods	M2S (103)				USPTO-M (74)	WildMol-M (10000)	IP5-M (1000)
	CXSMILES	Table		Markush	CXSMILES	CXSMILES	CXSMILES
	A	A	F1	A	A	A	A
Image only							
MolParser-Base	39	–	–	–	30	38.1‡	47.7
MolScribe	21†	–	–	–	7†	28.1	22.3
Multimodal							
GPT-5	3	8	24	0			
DeepSeek-OCR	0	–	–	–	0	1.9	0.0
MarkushGrapher-1	38†	**29**†	**65**†	10†	32†	–	–
MarkushGrapher-2 (Ours)	**56**	22	**65**	**13**	**55**	**48.0**	**53.7**

Table 2. **Markush Structure Recognition:** Comparison of our MarkushGrapher-2 model with existing models on CXSMILES and Substituent Table prediction. Models are evaluated on real-world benchmarks (M2S, UPSTO-M, WildMol-M, IP5-M). Accuracy A measures the percentage of correctly predicted samples, while F1 quantifies the similarity between predictions and ground truth, ranging from 0 (least similar) to 100 (most similar). † scores taken from [18], ‡ scores taken from [9].

Methods	WildMol (10000)	JPO (450)	UOB (5740)	USPTO (5719)
Image only				
MolParser-Base	**76.9**	**78.9**	91.8	93.0
MolScribe	66.4	76.2	87.4	**93.1**
DECIMER 2.7	56.0	64.0	88.3	59.9
MolGrapher	45.5	67.5	94.9	91.5
Multi-modal				
GPT-5		19.2		
DeepSeek-OCR	25.8	31.6	78.7	36.9
MarkushGrapher-1	-	-	-	-
MarkushGrapher-2 (Ours)	68.4	71.0	**96.6**	89.8

Table 3. **Molecular Structure Recognition**: Comparison of our MarkushGrapher-2 model with existing models on SMILES prediction. Models are evaluated on real-world benchmarks (WildMol, JPO, UOB, and USPTO). Accuracy measures the percentage of correctly predicted molecular structure. Scores for MolParser, MolScribe, DECIMER, and MolGrapher scores are taken from [9, 18].

Methods	M2S		USPTO-M		IP5-M	
	A	A_{InChIKey}	A	A_{InChIKey}	A	A_{InChIKey}
without OCR	4	39	3	51	15.4	51.3
with OCR	56	80	55	69	53.7	73.3

Table 4. **Effect of ChemicalOCR on Overall Model Performance**: Comparison of MarkushGrapher-2 performance with and without OCR predictions as input. The table shows the CXSMILES prediction accuracy A and structural backbone prediction accuracy A_{InChIKey}.

Methods	M2S		JPO	
	A	A_{InChIKey}	A	A_{InChIKey}
Fusion only	44	53	53.0	53.0
Adaptation and Fusion	50	68	61.5	61.5

Table 5. **Effect of Two-Phase Training**: Comparison of MarkushGrapher-2 performance with and without two-phase training. 'Fusion only' denotes the model trained in a single phase, while 'Adaptation and Fusion' refers to the two-phase training setup. Scores for CXSMILES prediction accuracy A and structural backbone prediction accuracy A_{InChIKey} are reported after 2 epochs for both configurations.

6. Conclusion

In this work, we present MarkushGrapher-2, a universal model for the recognition of both molecular structures and multi-modal Markush structures. Our approach introduces a dedicated ChemicalOCR module that extracts text from images, allowing the joint encoding of image, text, and layout modalities for end-to-end processing. To

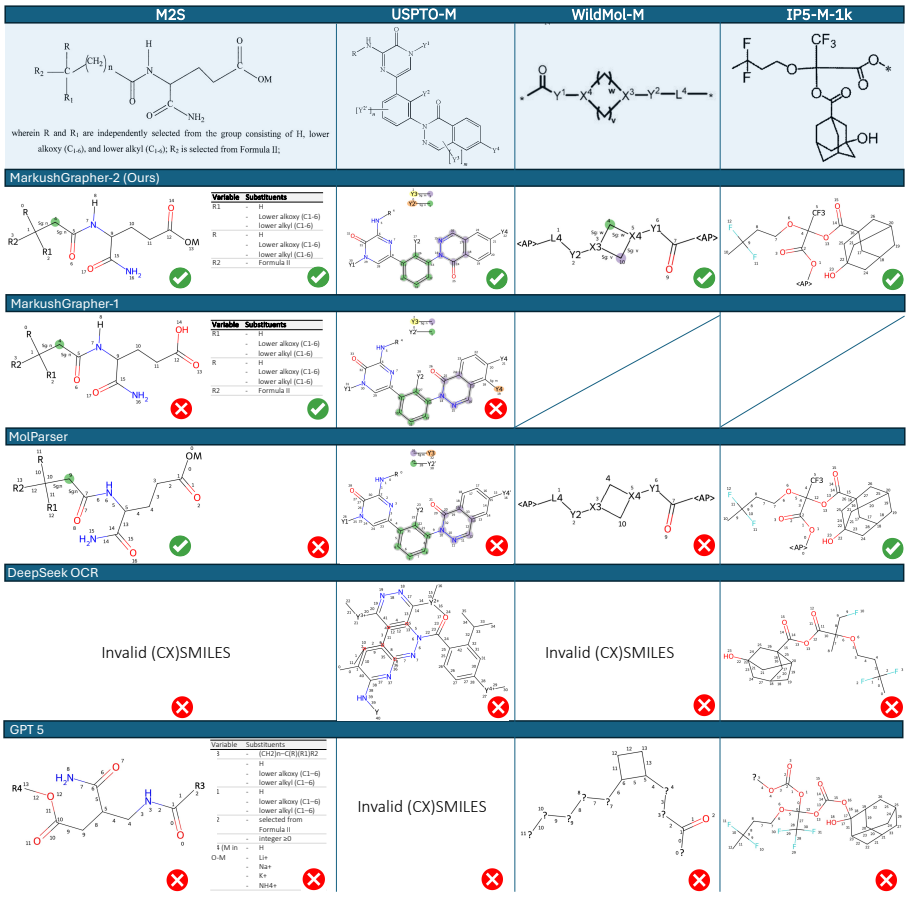

Figure 5. **Markush Structure Recognition - Qualitative Comparison**: Comparison of Markush structure predictions by five models MarkushGrapher-2 (Ours), MarkushGrapher-1, MolParser, DeepSeek OCR, and GPT-5 for an exemplary Markush structure from the benchmarks M2S, USPTO-M, WildMol-M, and IP5-M. Red labels indicate incorrect predictions, green labels indicate correct predictions.

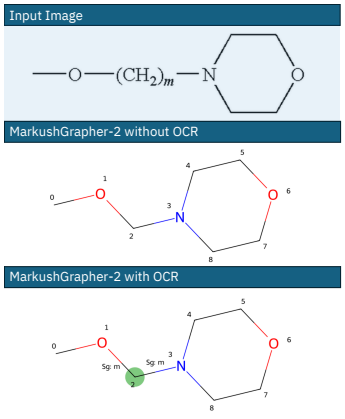

Figure 6. **Effect of OCR Predictions**: Comparison of MarkushGrapher-2 predictions with and without OCR input. The green circle indicates prediction of a frequency variation indicator (i.e., repeating Sg groups).

train MarkushGrapher-2, we employ a two-phase training strategy designed to optimally leverage pretrained features from a vision encoder, effectively fusing them for improved Markush structure recognition. To address the scarcity of training data, we developed a data generation pipeline for constructing real-world Markush samples from MOL files and accompanying USPTO images. In addition, we introduce IP5-M, a manually annotated benchmark dataset of real-world Markush structures from IP5 patent documents. Extensive experiments demonstrate that MarkushGrapher-2 substantially outperforms state-of-the-art models on Markush structure recognition, while remaining competitive on standard molecular structure recognition tasks. Our work bridges the gap between molecular and Markush recognition, offering a unified solution for large-scale automated extraction of chemical structures in documents.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*, 2025. 2

[2] 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] 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

[4] Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, Yue Zhang, Wenyu Lv, Kui Huang, Yichao Zhang, Jing Zhang, Jun Zhang, Yi Liu, Dianhai Yu, and Yanjun Ma. Paddleocr 3.0 technical report, 2025. 5

[5] 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. 2

[6] 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

[7] 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

[8] Xi Fang, Jiankun Wang, Xiaochen Cai, Shangqian Chen, Shuwen Yang, Haoyi Tao, Nan Wang, Lin Yao, Linfeng Zhang, and Guolin Ke. Molparser: End-to-end visual recognition of molecule structures in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 24528–24538, 2025. 2

[9] Xi Fang, Jiankun Wang, Xiaochen Cai, Shangqian Chen, Shuwen Yang, Haoyi Tao, Nan Wang, Lin Yao, Linfeng Zhang, and Guolin Ke. Molparser: End-to-end visual recognition of molecule structures in the wild, 2025. 3, 4, 5, 7

[10] Igor V Filippov and Marc C Nicklaus. Optical structure recognition software to recover chemical information: OSRA, an open source solution. *J. Chem. Inf. Model.*, 49(3): 740–743, 2009. 5

[11] 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

[12] Akio Fujiyoshi, Koji Nakagawa, and Masakazu Suzuki. Robust method of segmentation and recognition of chemical structure images in cheminfty. 2011. 5

[13] JaidedAI. EasyOCR: Ready-to-use OCR with 80+ supported languages and all popular writing scripts including latin, chinese, arabic, devanagari, cyrillic and etc, 2024. Available at: https://github.com/JaidedAI/EasyOCR. 5

[14] 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 2025 update. *Nucleic Acids Res.*, 53(D1): D1516–D1525, 2025. 4

[15] Ilia 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

[16] 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

[17] 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

[18] Lucas Morin, Valery Weber, Ahmed Nassar, Gerhard Ingmar Meijer, Luc Van Gool, Yawei Li, and Peter Staar. MarkushGrapher: Joint Visual and Textual Recognition of Markush Structures . In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14505–14515, Los Alamitos, CA, USA, 2025. IEEE Computer Society. 1, 2, 4, 5, 7

[19] 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

[20] 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

[21] 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, 4, 5

[22] 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

[23] 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

[24] Kohulan Rajan, Viktor Weißenborn, Laurin Lederer, Achim Zielesny, and Christoph Steinbeck. Marcus: molecular annotation and recognition for curating unravelled structures. *Digital Discovery*, 4:3137–3148, 2025. 1

[25] Yaorui Shi, Sihang Li, Taiyan Zhang, Xi Fang, Jiankun Wang, Zhiyuan Liu, Guojiang Zhao, Zhengdan Zhu, Zhifeng Gao, Renxin Zhong, Linfeng Zhang, Guolin Ke, Weinan E, Hengxing Cai, and Xiang Wang. Intelligent system for automated molecular patent infringement assessment. 2025. 2

[26] Edlyn S. Simmons. Markush structure searching over the years. *World Patent Information*, 25(3):195–202, 2003. 2

[27] Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, Siyang Qin, Reeve Ingle, Emanuele Bugliarello, Sahar Kazemzadeh, Thomas Mesnard, Ibrahim Alabdulmohsin, Lucas Beyer, and Xiaohua Zhai. PaliGemma 2: A Family of Versatile VLMs for Transfer. *arXiv preprint arXiv:2412.03555*, 2024. 2

[28] 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

[29] 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

[30] 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. 3, 4

[31] 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

[32] 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

[33] Haoran Wei, Yaofeng Sun, and Yukun Li. Deepseek-ocr: Contexts optical compression, 2025. 2

[34] 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

[35] 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