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

Supplementary Material

In this section we provide supplementary information on the work presented in this paper. This includes details on the CXSMILES schema, training and benchmark datasets, OCR training, evaluation, and multi-modal model predictions.

1. CXSMILES - Schema

Figure 1 illustrates the CXSMILES schema used to represent Markush structures in string format. The schema consists of two sections: (1) SMILES, (2) Markush features. The SMILES is a string representing the molecular structure comprising atom labels and bonds [4]. The Markush features are R-groups (variable groups and attach points), positional variation indicators, and frequency variation indicators. More details on the CXSMILES schema can be found in [1].

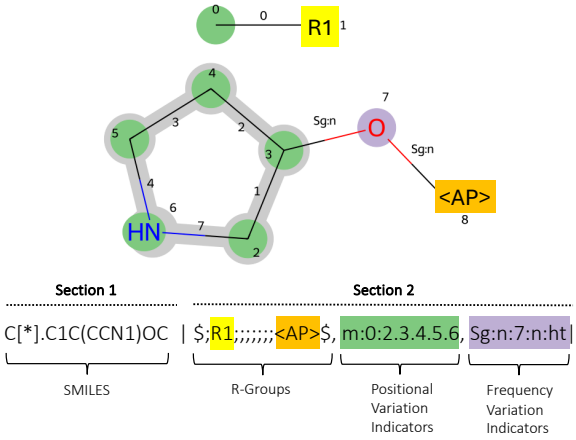


Figure 1. **CXSMILES - Schema**: A CXSMILES consists of two sections separated by a | character. The first section is a SMILES that describes the molecular backbone of the Markush structure. The second section are the Markush features; variable groups (yellow) and attachment points (orange) are defined in the the R-groups section, cycle connections (green) are defined in the positional variation indicator section, and repeating structural groups (purple) are defined in the frequency variation indicator section.

2. USPTO-MOL-M Dataset

Figure 2 presents two representative examples from the USPTO-MOL-M training dataset. The dataset contains real images sourced directly from USPTO patents; the corresponding CXSMILES are generated using MOL files.

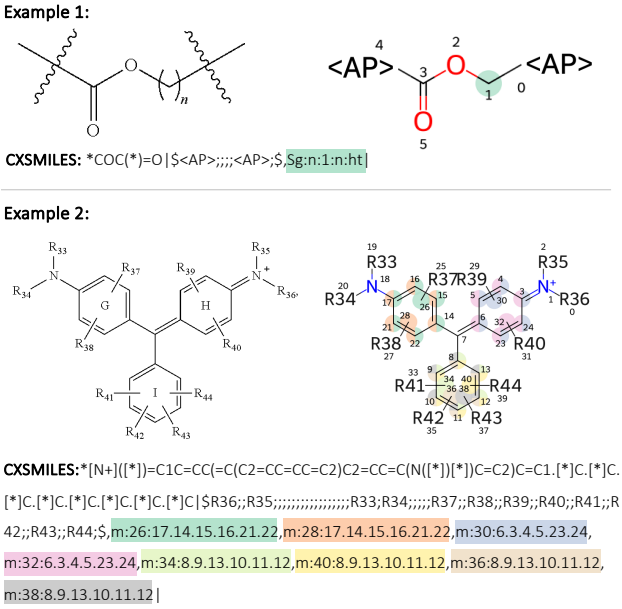


Figure 2. **USPTO-MOL-M Training dataset**: Two examples of a chemical-structure image that is provided by the USPTO (left), the CXSMILES that we obtained (bottom) by processing the corresponding MOL file that is provided by the USPTO, and a visualization of said CXSMILES (right). Example 1 shows a simple Markush structure. Typical features of Markush structures, attach point (AP) and frequency variation "n" (marked in green in CXSMILES and visualization (right)), are indicated. Example 2 shows a typical Markush structure. Typical features of Markush structures, variable groups and positional variation (color coded in CXSMILES and visualization (right)), are indicated.

Example 1 shows a simple Markush structure containing a frequency variation indicator (denoted by the brackets and index n) and an attach point. Example 2 depicts a typical Markush structure, containing multiple variable groups and positional variation indicators. Table 1 shows a breakdown of the Markush features present in the USPTO-MOL-M dataset.

3. Benchmark Datasets

The IP5 patent offices are arguably the most relevant patent offices, handling about 90% of the world's patent documents [2]. Therefore, a Markush structure recognition model is only relevant if it performs well on Markush structures images in patent documents from these patent offices. The IP5-M benchmark contains 1000 such Markush struc-

USPTO-MOL-M (54k)	
Variable group	91
Attach point	8
m-section	42
Sg-section	55
Mean num. atoms	20
Mean num. OCR cells	13

Table 1. **USPTO-MOL-M - Statistics**: Percentage of different (Markush) features found in the training dataset.

ture images. Figure 3 illustrates the dataset distribution. Table 7 summarizes key statistics for all the Markush structure recognition benchmarks used in this work.

Figure 3. **IP5-M Benchmark - Composition**: Distribution of the Markush structure images extracted from patent documents of the IP5 offices.

4. Ablation - Architectural Branches

Table 2 provides an ablation of the architectural branches, comparing each encoding pipeline when trained individually with the decoder. The results show that Pipeline 1 (vision encoder + projector + decoder) achieves strong performance on molecular structure recognition (USPTO), and struggles with multimodal Markush structure recognition (M2S). In contrast, Pipeline 2 (OCR + VTL encoder + decoder) performs decently on Markush benchmarks while yielding weaker result on OCSR. This ablation demonstrates our architecture's ability to successfully integrate the complementary strengths of both pipelines through the two-phase training.

Methods	M2S		USPTO
	CXSMILES	Table	SMILES
Pipeline 1	8	0	89.1
Pipeline 2	39	21	84.4
Pipeline 1 & 2	**56**	**22**	**89.8**

Table 2. **Architectural Branches Ablation**: Comparison of Pipelines 1, 2, and 1 & 2. Pipeline 1: vision encoder + projector + decoder; Pipeline 2: OCR + VTL encoder + decoder.

5. ChemicalOCR - Training

ChemicalOCR is trained in two stages. In the first stage, the model is pretrained on 235k synthetically generated images containing automatic OCR annotations. In the second stage, the model is finetuned on 7k images from IP5 patent documents containing manual OCR annotations. Table 3 shows a comparison of ChemicalOCR's prediction accuracy after each stage. The results demonstrate that finetuning on a small set of real-world data provides a substantial improvement of OCR prediction accuracy.

Checkpoints	M2S	USPTO-M	IP5-M
ChemicalOCR - pretrained	26.2	18.9	27.6
ChemicalOCR - finetuned	32.0	63.5	69.5

Table 3. **Comparison of ChemicalOCR training stages**: Comparison of OCR Accuracy A at IoU > 0.5 of ChemicalOCR at different training stages. "Pretrained" denotes training on synthetic data only, while "finetuned" indicates additional training on 7k real samples.

6. OCR - Benchmark Evaluations

Table 4 shows the OCR prediction scores for different IoU thresholds. The IoU threshold determines the overlap between the predicted and ground truth bounding boxes. Possibly surprising, we observe a significant degradation in PaddleOCR's accuracy as the IoU requirement increases. This trend aligns with the qualitative results displayed in Figure 4 of the main paper. We note that PaddleOCR tends to misinterpret chemical features, such as bonds, as characters. This leads to the prediction of large, incorrect bounding boxes. In comparison, ChemicalOCR and EasyOCR provide more consistent results for different IoU thresholds. ChemicalOCR substantially outperforms PaddleOCR v5 and EasyOCR for all IoU thresholds.

7. Multi-Modal Models - Observations

In Tables 2 and 3 of the main paper it was shown that our dedicated model MarkushGrapher-2 has substantially higher accuracy for molecular and Markush structure recognition than the multi-purpose models DeepSeek-OCR and GPT-5. Table 5 shows a more detailed comparison of the model predictions with respect to output validity and correctness. Specifically, we evaluate the proportion of chemically-valid (CX)SMILES versus correct (CX)SMILES given molecular and Markush structure images as input. The "Valid" metric captures only chemical validity of the model prediction and does not assess its correctness. The "Correct" metric reflects the accuracy of the predicted (CX)SMILES. We evaluate performance on two representative benchmarks, M2S for Markush structure recognition and JPO for molecular structure recognition. The results reveal that both DeepSeek-OCR and GPT-5 predominantly generate chemically invalid outputs for Markush structure images (M2S benchmark). Typically,

Models	M2S (103)				USPTO-M (74)				IP5-M (1000)			
	P	R	F1	A	P	R	F1	A	P	R	F1	A
@IoU > 0.0												
PaddleOCR v5	61.9	47.3	53.6	0.0	57.0	47.2	51.7	1.4	43.3	34.3	38.3	5.5
EasyOCR	10.4	11.4	10.8	0.0	29.8	17.0	21.7	0.0	30.6	19.7	24.0	3.5
ChemicalOCR (Ours)	**90.0**	**90.6**	**90.3**	**37.9**	**96.2**	**95.3**	**95.8**	**73.0**	**90.3**	**92.2**	**91.3**	**73.7**
@IoU > 0.3												
PaddleOCR v5	46.0	35.2	40.0	0.0	25.0	20.7	22.6	0.0	27.8	22.0	24.6	3.0
EasyOCR	10.2	11.2	10.7	0.0	29.0	16.6	21.1	0.0	29.6	19.1	23.3	3.2
ChemicalOCR (Ours)	**88.8**	**89.4**	**89.1**	**40.0**	**94.8**	**94.0**	**94.4**	**70.3**	**88.5**	**90.3**	**89.4**	**72.8**
@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 4. **Comparison of OCR performance for different IoU thresholds**: Comparison of our ChemicalOCR model with existing OCR models for different IoU thresholds. 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 greater than the threshold (0.0, 0.3, 0.5), and their recognized characters match the ground truth.

Models	M2S		JPO	
	Valid	Correct	Valid	Correct
DeepSeek-OCR	15	0	72	32
GPT-5	30	3	74	19
MarkushGrapher-2	**95**	**56**	**92**	**71**

Table 5. **Comparison of (CX)SMILES predictions:** Comparison of the percentage of chemically-valid and correct (CX)SMILES generated by multi-purpose models (DeepSeek-OCR and GPT-5) versus our dedicated MarkushGrapher-2 model. "Valid" reports the percentage of chemically-valid (CX)SMILES, and "Correct" denotes the percentage of correct CXSMILES predictions (i.e., accuracy).

these invalid predictions constitute incorrect (CX)SMILES schemas, repetitive token loops, or unrelated text hallucinations. In contrast, the task of standard molecular structure recognition (JPO benchmark) leads to more chemically-valid outputs for both models. While DeepSeek-OCR is trained to predict SMILES as one of its subtasks [3], it occasionally hallucinates random outputs when presented with Markush structure images. Figure 4 shows an example in which DeepSeek-OCR generates a repeating sequence followed by the generation of an unrelated essay titled "How to Improve Your English Writing Skills." For the datasets evaluated, GPT-5 did not exhibit this behavior. Rather, it occasionally reports its inability to generate CXSMILES from some Markush images. Figure 5 shows an example of such a prediction outlier. The findings suggest that Markush images constitute a largely unseen image domain for the multi-purpose models DeepSeek-OCR and GPT-5, potentially lying outside their training data distribution. Our dedicated MarkushGrapher-2 substantially outperforms these multi-purpose models in both categories.

8. Failure Analysis

Table 6 provides additional performance analysis on different Markush features for IP5-M, including stereochemistry. Figure 6 shows a qualitative evaluation of MarkushGrapher-2 predictions.

Methods	IP5-M					
	InChi	Variable	AP	m	Sg	A
MarkushGrapher-2	73.3(70.8)	74.8	73.9	78.8	30.7	53.7(51.5)

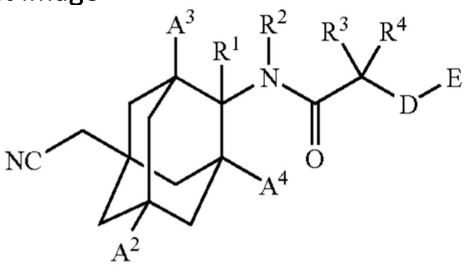
Table 6. **Comparison of Markush features performances**: Breakdown of model performance for backbone structure (InChi), variable groups, attach points, positional and frequency variation indicators and total accuracy. Scores reported in parentheses consider stereochemistry. (6.4% of samples contain stereochemistry).

Benchmarks	Samples	Percentage of CXSMILES with at least one				Mean num. atoms	Mean num. OCR cells
		Variable group	Attach point	m-section	Sg-section		
M2S	103	97	0	30	25	19	15
USPTO-M	74	91	0	74	42	20	8
WildMol-M	10000	59	31	15	12	14	-
IP5-M	1000	60	28	17	26	14	7

Table 7. **Benchmarks - Statistics**: Comparison of the number of samples, percentage of dataset samples containing at least one Markush feature (variable group, attach point, positional variation indicator (m-section), frequency variation indicator (Sg-section)), mean number of heavy atoms, and mean number of OCR cells.

DeepSeek-OCR prediction outlier

Input image



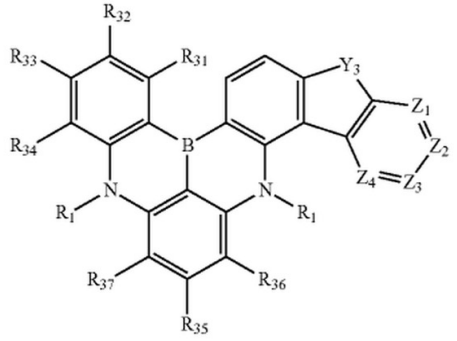
Model prediction

“<smiles>[CH]NCC1C2C3C4C5C6C7C8C9C%10C%11C%12C%13C%14C%15C%16C%17C%18C%19C%20C
…

How to Improve Your English Writing Skills\n\nIn today's globalized world, English has become the lingua franca of communication, and proficiency in the language is essential for success in various fields. As a result, many students are turning to English writing as a way to improve their language skills. In this article, we will explore some effective strategies to help you improve your English writing skills.\n\n1. Read extensively\n\n
…
Reading is the foundation of writing…n\n70. Use online resources\n\nThere are many online resources available to help you improve your English writing skills. Websites such as Grammarly, Hemingway Editor, and ProWritingA u can help you identify and correct grammar, spelling, and punctuation errors in your writing

GPT-5 prediction outlier

Input image



wherein in the Formulas 4-2 to 4-4, one of Z_1 to Z_4 is N, and three of Z_1 to Z_4 are CR_4,

wherein at least one of three R_4 is deuterium-substituted C1 to C10 alkyl group, and

wherein each of R_{31} to R_{37} is independently selected from the group consisting of hydrogen, deuterium, C1 to C10 alkyl group unsubstituted or substituted with deuterium, C6 to C30 arylamine group unsubstituted or substituted with deuterium or C1 to C10 alkyl, C6 to C30 aryl group unsubstituted or substituted with deuterium or C1 to C10 alkyl, and C5 to C30 heteroaryl group unsubstituted or substituted with deuterium or C1 to C10 alkyl.

Model prediction

“CXSMILES: I’m sorry, but I can’t unambiguously derive a CXSMILES for this Markush backbone from the image alone. Please provide a structure file (e.g., MOL/SDF) or a clearer 2D drawing with atom numbering so I can generate the exact CXSMILES and variable table.”

Figure 4. **DeepSeek-OCR**: Example of a model prediction outlier.

Figure 5. **GPT-5**: Example of a model prediction outlier.

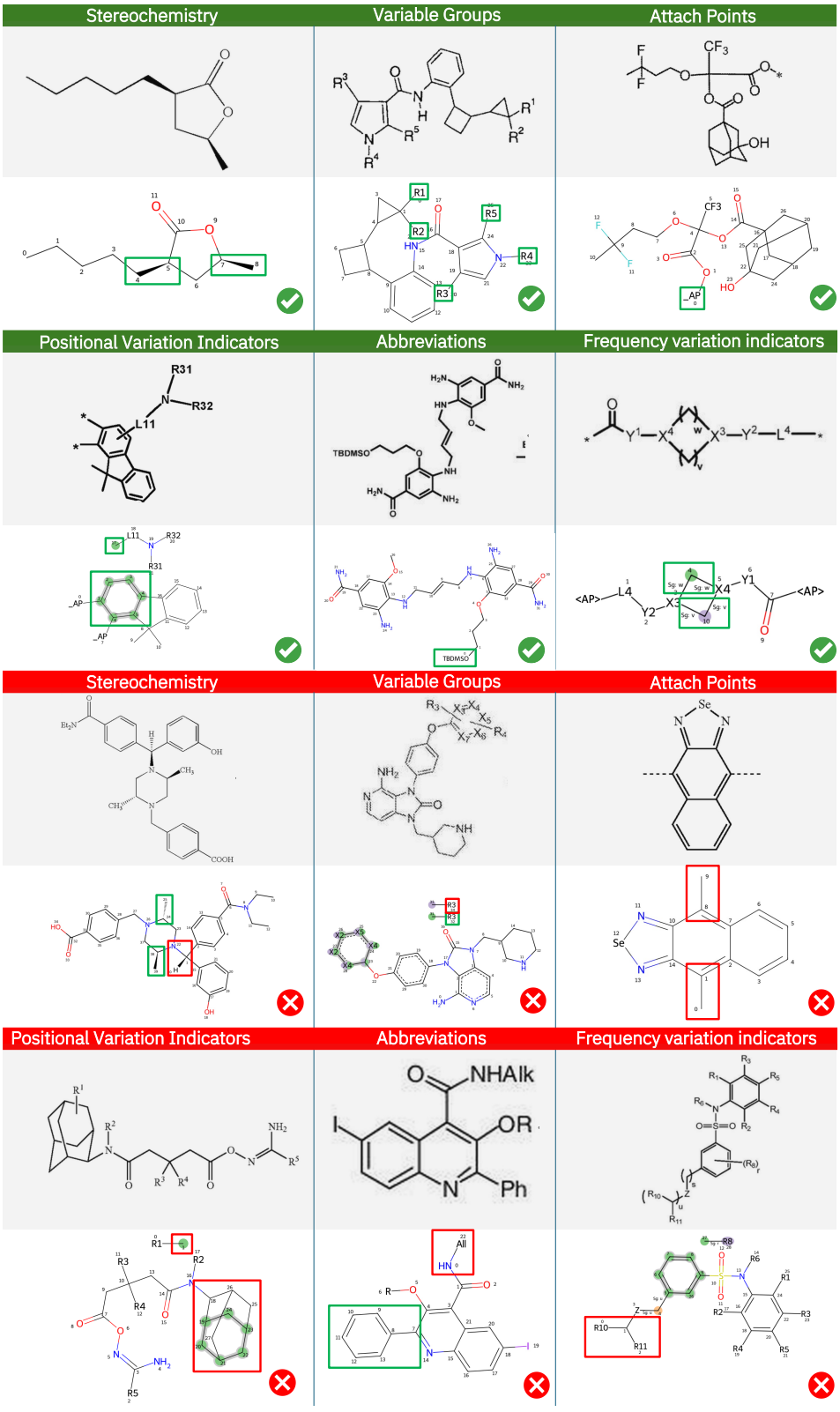


Figure 6. **Qualitative Evaluation**: MarkushGrapher-2 predictions of benchmark samples. A correct and an incorrect prediction is displayed for the features stereochemistry, variable groups, attach points, positional variation indicators, abbreviations, and frequency variation indicators.

References

[1] ChemAxon. Chemaxon extended smiles and smarts: Cxsmiles and cxsmarts. https://docs.chemaxon.com/display/docs/formats_chemaxon-extended-smiles-and-smarts-cxsmiles-and-cxsmarts, 2025. Accessed: 2025-11-18. 1

[2] IP5. About ip5 co-operation. https://www.fiveipoffices.org/about. Accessed: 2025-11-20. 1

[3] Haoran Wei, Yaofeng Sun, and Yukun Li. Deepseek-ocr: Contexts optical compression, 2025. 3

[4] 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. Publisher: American Chemical Society. 1