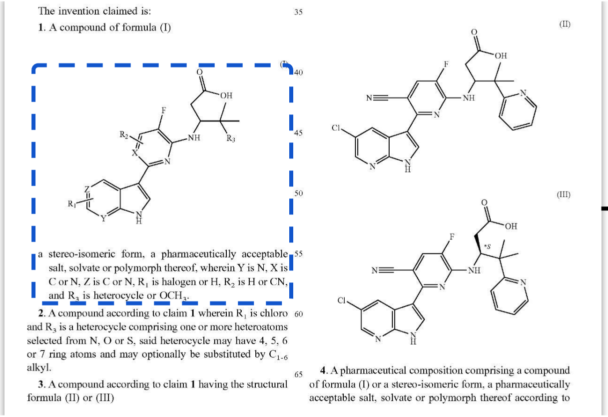
MarkushGrapher: Joint Visual and Textual Recognition of Markush Structures

Lucas Morin[1, 2], Valéry Weber[1], Ahmed Nassar[1], Gerhard Ingmar Meijer[1], Luc Van Gool[2, 3], Yawei Li[2], Peter Staar[1]

1 IBM Research 2 ETH zürich 3 INSAIT

I. Introduction: Multimodal Markush Structure Recognition

Document


The invention claimed is:

1. A compound of formula (I)

(I)

a stereo-isomeric form, a pharmaceutically acceptable salt, solvate or polymorph thereof, wherein Y is N, X is C or N, Z is C or N, R_1 is halogen or H, R_2 is H or CN, and R_3 is heterocycle or OCH_3.

2. A compound according to claim 1 wherein R_1 is chloro and R_3 is a heterocycle comprising one or more heteroatoms selected from N, O or S, said heterocycle may have 4, 5, 6 or 7 ring atoms and may optionally be substituted by C_{1-6} alkyl.

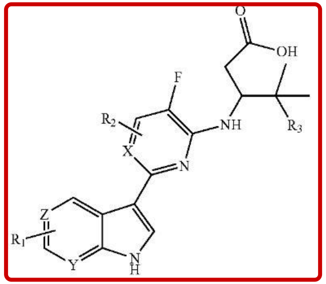
3. A compound according to claim 1 having the structural formula (II) or (III)

(II)

(III)

4. A pharmaceutical composition comprising a compound of formula (I) or a stereo-isomeric form, a pharmaceutically acceptable salt, solvate or polymorph thereof according to

Visual Component

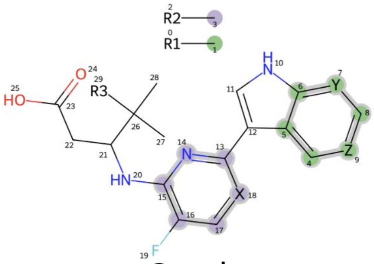


a stereo-isomeric form, a pharmaceutically acceptable salt, solvate or polymorph thereof, wherein Y is N, X is C or N, Z is C or N, R_1 is halogen or H, R_2 is H or CN, and R_3 is heterocycle or OCH_3.

Textual Component

MarkushGrapher

Markush Structure



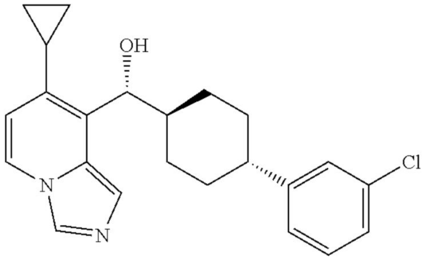
Graph

Y	N	
X	C	N
Z	C	N
R1	halogen	H
R2	H	CN
R3	heterocycle	OCH3

Table

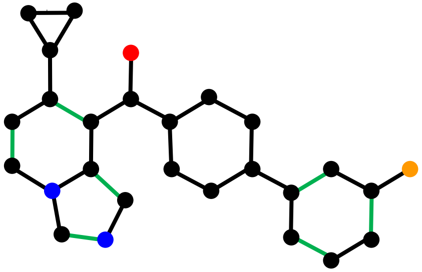
I. Introduction: Multimodal Markush Structure Recognition

Molecule Image



OCSR: Optical Chemical Structure Recognition (MolGrapher [1])

Molecular Graph



Text Segment

Triphosgene (0.556g, 1.87 mmol) was added over 20 mins to a solution of 8-hydroxy-1,6,naphthyridine-7-carbo xylic acid (0.89g, 4.68 mmol) and diis-opropylethylamine (3.26 ml, 18.7 mmol) in DMF (22 ml) at 0 °C.

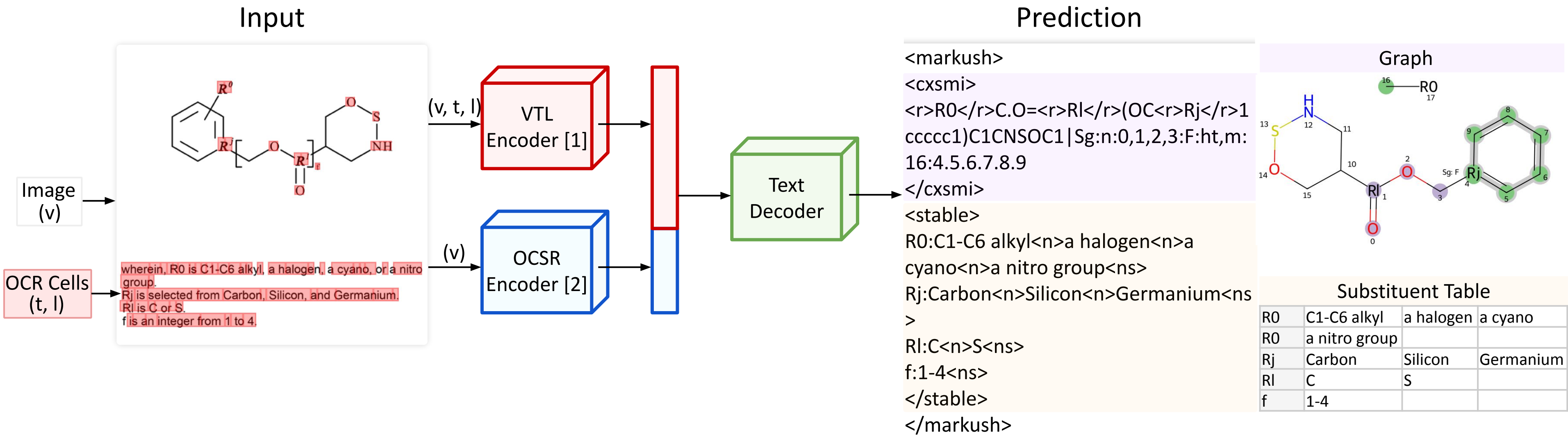
CNER: Chemical Named Entity Recognition (OSCAR4 [2])

Chemical Entities

Triphosgene (0.556g, 1.87 mmol) was added over 20 mins to a solution of 8-hydroxy-1,6,naphthyridine-7-carbo xylic acid (0.89g, 4.68 mmol) and diis-opropylethylamine (3.26 ml, 18.7 mmol) in DMF (22 ml) at 0 °C.

[1] Lucas Morin, *et al.*, MolGrapher: Graph-based Visual Recognition of Chemical Structures. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
[2] Jessop, *et al.*, OSCAR4: a flexible architecture for chemical text-mining. Journal of Cheminformatics (2011)

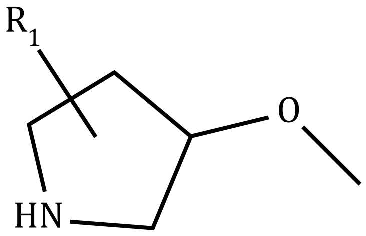
II. Method: MarkushGrapher

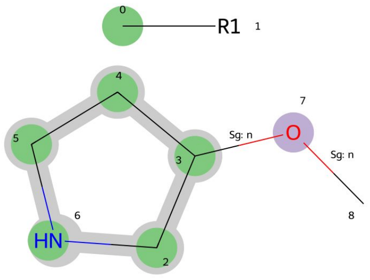


[1] Zineng Tang, *et al.*, Unifying Vision, Text, and Layout for Universal Document Processing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
[2] Yujie Qian, *et al.*, MolScribe: Robust Molecular Structure Recognition with Image-to-Graph Generation. Journal of Chemical Information and Modeling (2023)

II. Method: Optimized Markush Structure Representation

Markush Structure Backbone





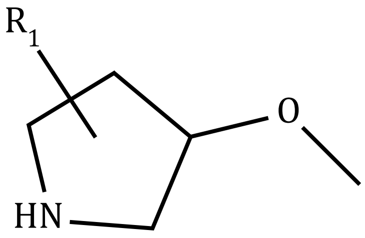
Sequence Representation Optimization

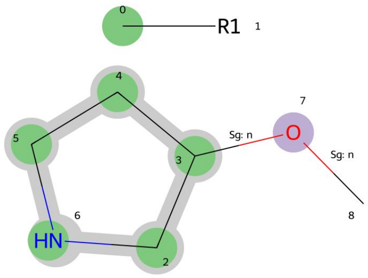
1. C[*].C1C(CCN1)OC |$;R1;;;;;;;;;$,m:0:2.3.4.5.6,Sg:n:7:n:ht| CXSMILES [1]

[1] ChemAxon Extended SMILES (CXSMILES) Documentation. https://docs.chemaxon.com/display/docs/formats_chemaxon-extended-smiles-and-smarts-cxsmiles-and-cxsmarts. Accessed: 2024.

II. Method: Optimized Markush Structure Representation

Markush Structure Backbone





Sequence Representation Optimization

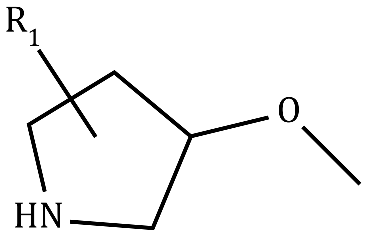
1. C[*].C1C(CCN1)OC |$;R1;;;;;;;;;$,m:0:2.3.4.5.6,Sg:n:7:n:ht| — CXSMILES [1]

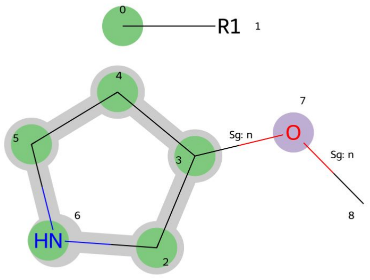
2. C<r>R1</r>.C1C(CCN1)OC|m:0:2.3.4.5.6,Sg:n:7:n:ht — Compact the CXSMILES by repositioning variable groups into the SMILES sequence and removing unnecessary characters

[1] ChemAxon Extended SMILES (CXSMILES) Documentation. https://docs.chemaxon.com/display/docs/formats_chemaxon-extended-smiles-and-smarts-cxsmiles-and-cxsmarts. Accessed: 2024.

II. Method: Optimized Markush Structure Representation

Markush Structure Backbone





Sequence Representation Optimization

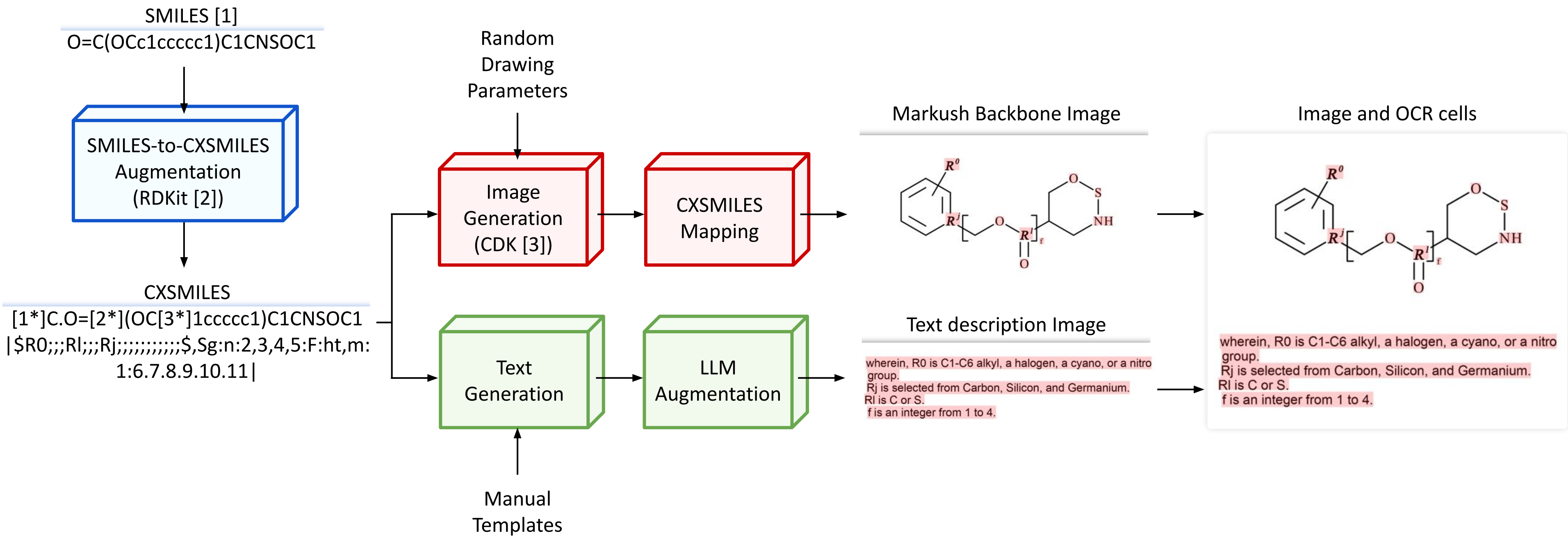
1. C[*].C1C(CCN1)OC |$;R1;;;;;;;;;$,m:0:2.3.4.5.6,Sg:n:7:n:ht| — CXSMILES [1]

2. C<r>R1</r>.C1C(CCN1)OC|m:0:2.3.4.5.6,Sg:n:7:n:ht — Compact the CXSMILES by repositioning variable groups into the SMILES sequence and removing unnecessary characters

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 — Append atom indices inline, and encode atoms and bonds using special tokens

[1] ChemAxon Extended SMILES (CXSMILES) Documentation. https://docs.chemaxon.com/display/docs/formats_chemaxon-extended-smiles-and-smarts-cxsmiles-and-cxsmarts. Accessed: 2024.

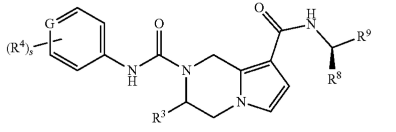
II. Method: Synthetic Data Generation



[1] Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. Journal of Chemical Information and Computer Sciences (1988)
[2] Landrum, G. et al. RDKit: Open-Source Cheminformatics Software. http: //www.rdkit.org/ (2006)
[3] Willighagen, E. L., *et al.*, The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. Journal of Cheminformatics (2017)

II. Method: M2S and USPTO-Markush Benchmarks

M2S

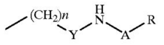


wherein:

G is C or N;

R^8 is methyl; and

R^9 is methyl or ethyl, wherein the methyl or ethyl is optionally further substituted by one or more halogen.



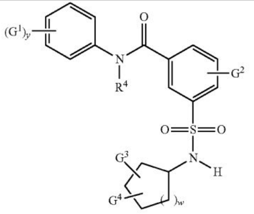
in which

n represents 1, 2, 3, 4, 5, 6, 7, 8 or 9

Y, A represent C═O or —SO_2, and

R represents an optionally substituted alkyl or aryl radical,

and at least one substituent selected from the group consisting of



or pharmaceutically acceptable salts thereof;

wherein G^1 is independently selected at each occurrence from CH_3, OCH_3, halo, CF_3, CCl_3, CH_2Cl, CCl_2H, CF_2H, CH_2F, and CF_3;

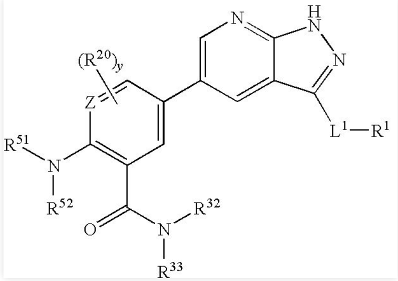
G^2 is H, C_1-C_4 alkyl, or halo;

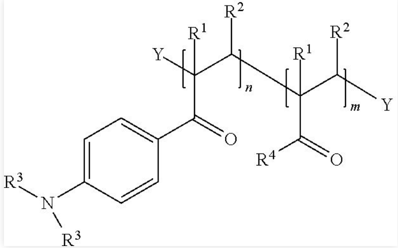
G^3 is OH, CH_2OH, or CH_2CH_2OH;

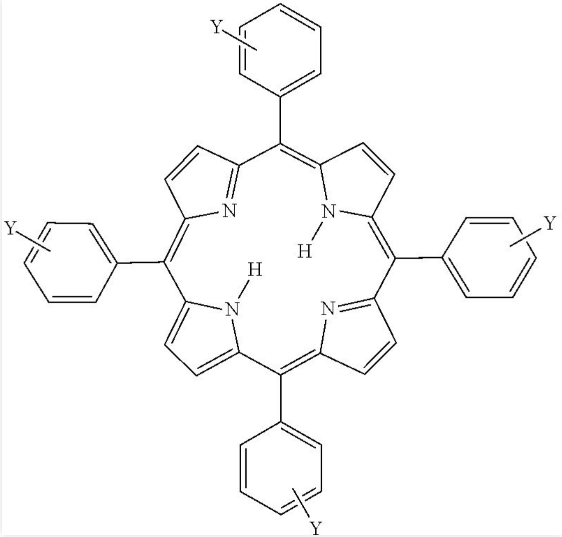
G^4 is H, OH, halo, C_1-C_6 alkyl, C_1-C_6 alkyl-OH, —C_1-C_6 chloroalkyl, —C_1-C_6 dichloroalkyl, —C_1-C_6 trichloroalkyl, —C_1-C_6 fluoroalkyl, —C_1-C_6 difluoroalkyl, —C_1-C_6 trifluoroalkyl, or phenyl, wherein the phenyl group is optionally independently substituted with 1-5 substituents selected from halo, —C_1-C_6 alkyl, and —C_1-C_6 alkoxy; and

y is 1, 2, or 3.

USPTO-Markush

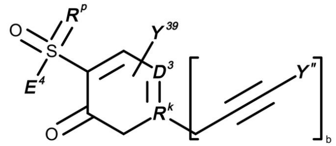






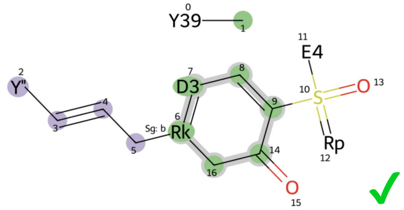
III. Results: Qualitative Examples

1. Synthetic



wherein
Y39 represents H11C5 and C21H43, Rp represents N-aryl, or S, D3 is selected from the group consisting of CH, or N, b is 18, Rk is N,N',N'-Carbohydrazide, Y" is independently selected from C1-C4alkoxy heteroaryl or a group represented by the formula (2), E4 is MgH, Rb, CH3, and NH2.

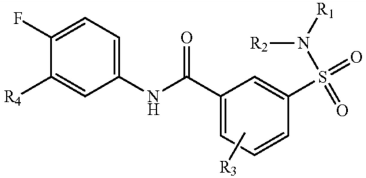
1. Prediction



Y39	H11C5	C21H43		
Rp	N-aryl	S		
D3	CH	N		
b	18			
Rk	N,N',N'-Carbohydrazide			
Y''	C1-C4alkoxy heteroaryl			
Y''	a group represented by the formula (2)			
E4	MgH	Rb	CH3	NH2

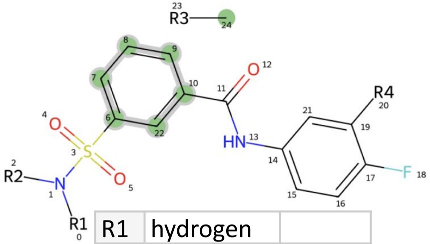
✓

2. M2S



or a stereoisomer or tautomeric form thereof, wherein:
R_1 represents hydrogen;
R_2 represents C_1-C_8alkyl substituted with one or more R_5,
R_3 represents Hydrogen or methyl;
R_4 represents methyl;

2. Prediction



R1	hydrogen	
R2	C1-C8 alkyl	
R3	Hydrogen	methyl
R4	methyl	

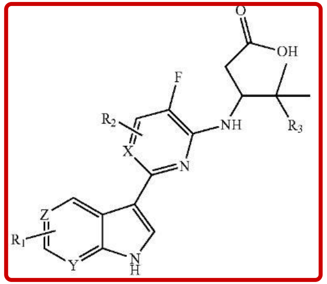
✓

III. Results: Comparison with State-of-the-art

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	7	35	-	-	-	3	25	-	-	-	7	40
MolScribe	11	47	-	-	-	21	73	-	-	-	7	**97**
Multi-modal												
Pixtral-12B-2409	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	0	6	10	46	0	4	11	19	49	0	-	-
Uni-Smart	-	-	-	-	-	0	46	0	1	0	-	-
MarkushGrapher (Ours)	**65**	**96**	**84**	**96**	**57**	**38**	**76**	**29**	**65**	**10**	**32**	74

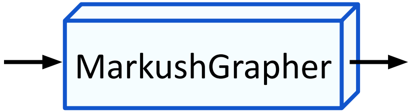
IV. Conclusion

Visual Definition

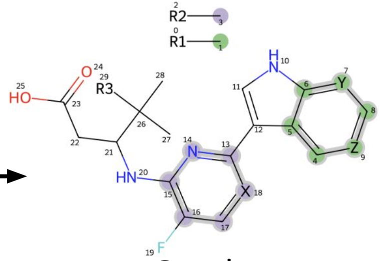


a stereo-isomeric form, a pharmaceutically acceptable salt, solvate or polymorph thereof, wherein Y is N, X is C or N, Z is C or N, R_1 is halogen or H, R_2 is H or CN, and R_3 is heterocycle or OCH_3.

Textual Definition



Markush Structure



Graph

Y	N	
X	C	N
Z	C	N
R1	halogen	H
R2	H	CN
R3	heterocycle	OCH3

Table

Code and datasets: https://github.com/DS4SD/MarkushGrapher

Contributions:

1. MarkushGrapher, a model combining a VTL encoder and an OCSR encoder, that significantly outperform SOTA for multi-modal Markush structure recognition.
2. Synthetic data generation pipeline producing diverse and realistic results.
3. M2S, benchmark of manually-annotated multi-modal Markush structures from patent documents.