THRUST 1

TRANSLATION BETWEEN MOLECULES AND NATURAL LANGUAGE

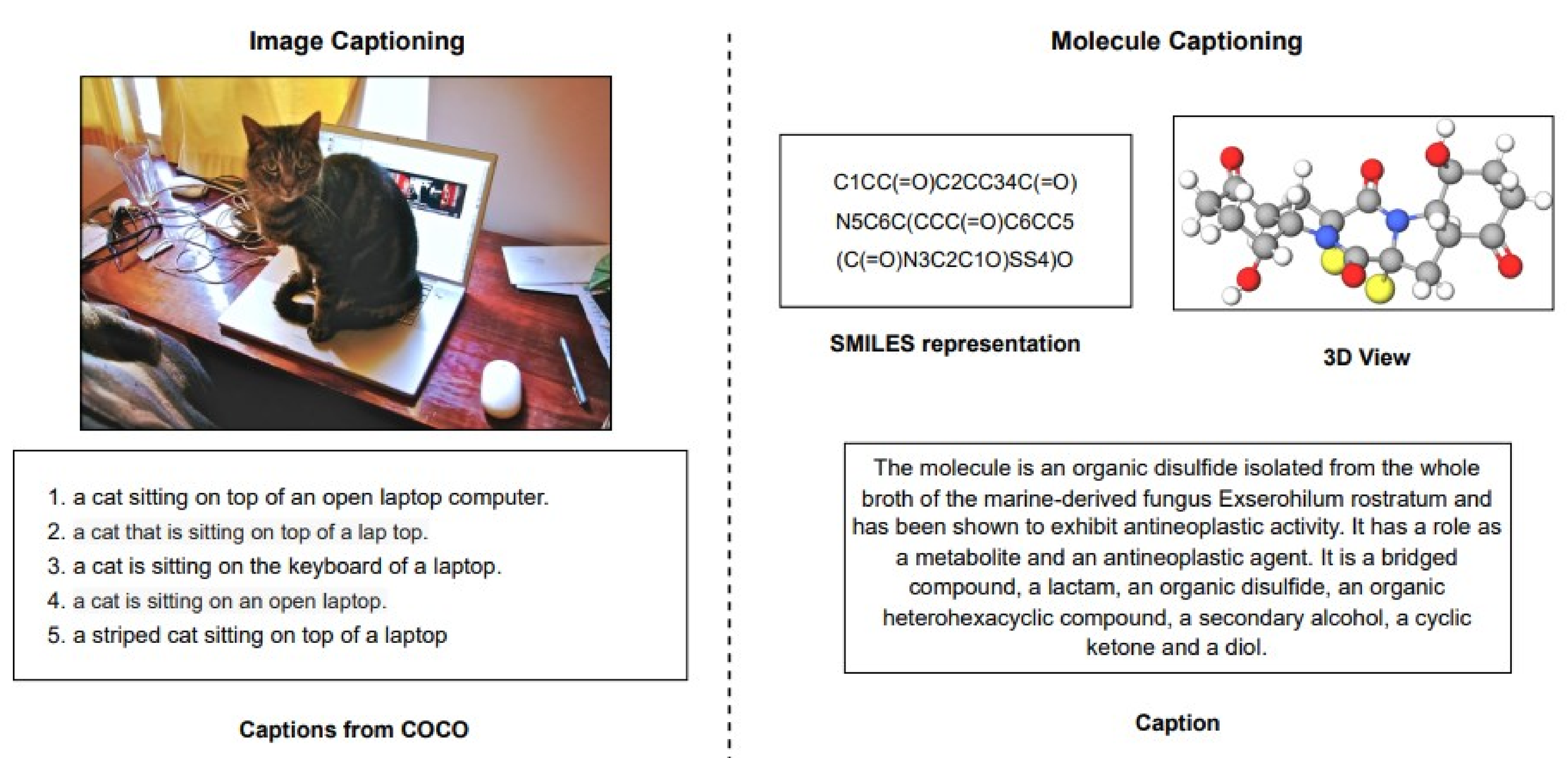
CARL EDWARDS*[1], TUAN LAI*[1,2], KEVIN ROS[1], GARRETT HONKE[2], KYUNGHYUN CHO[3], HENG JI[1]
[1]UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN, [2]X, THE MOONSHOT FACTORY, [3]NEW YORK UNIVERSITY

INTRODUCTION

- There has been a lot of recent success in combining language and images for enabling semantic control of images.
    - Can this be extended to molecules?
- There are some inherent difficulties:
    1. Creating annotations requires significant domain expertise
    2. Thus, it is more difficult to acquire large datasets
    3. The same molecule can be described many ways
    4. Existing evaluation measures for sequence generation are often inadequate
- To address these issues, we propose **MolT5**, a self-supervised learning framework inspired by recent progress one multilingual models, which is pretrained on single-modal data before task-specific finetuning.

TASK DEFINITION

- We propose two new tasks:
    1. molecule captioning, where a description is generated for a given molecule
    2. text-based de novo molecule generation, where a molecule is generated to match a given text description.

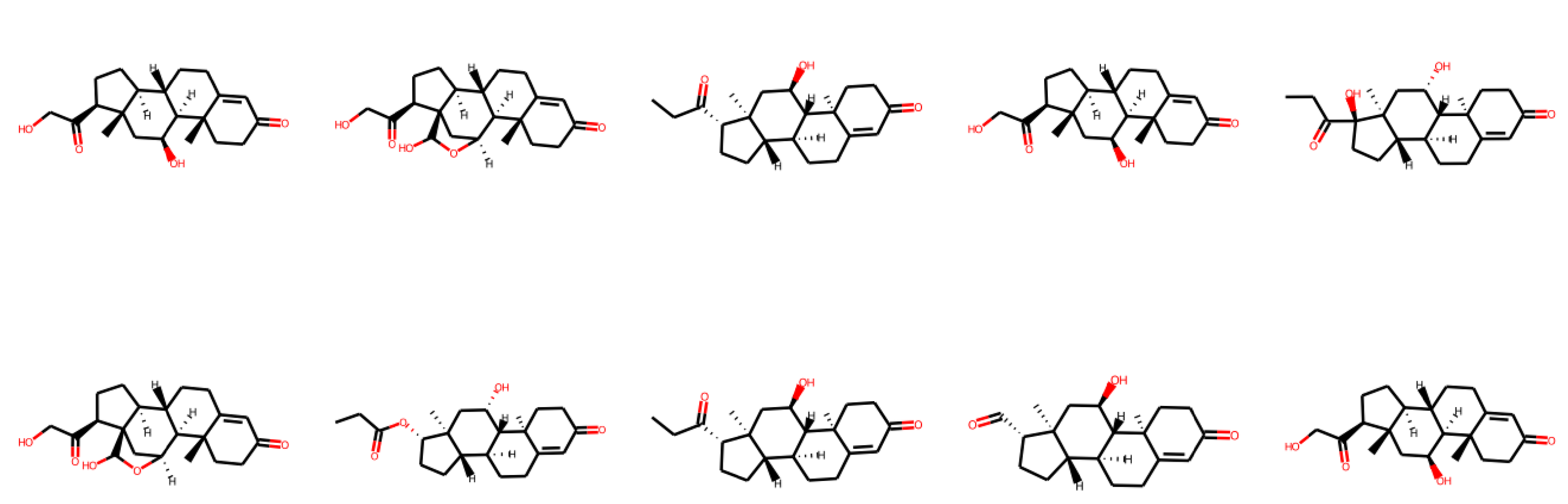


DATA

- We used CheBI-20, a dataset consisting of 33,010 compound-description pairs

PROBING TESTS

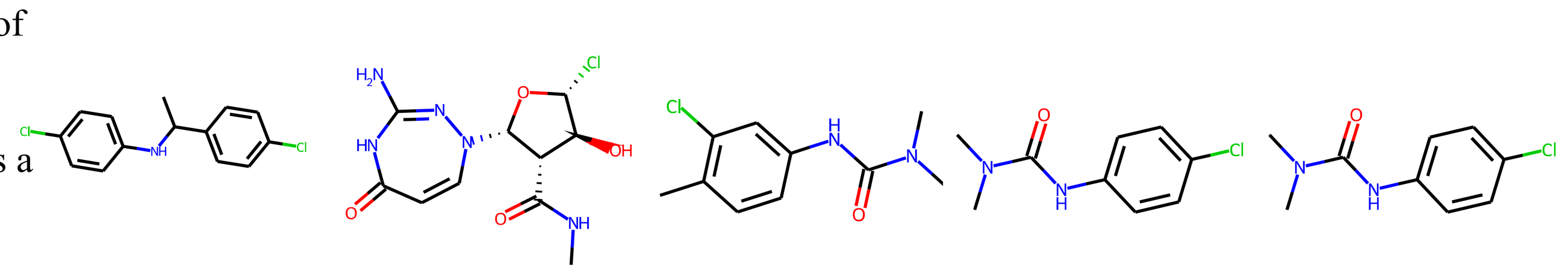
- We probe the model with different properties. Below, we input "The molecule is a corticosteroid."



MODEL OUTPUTS

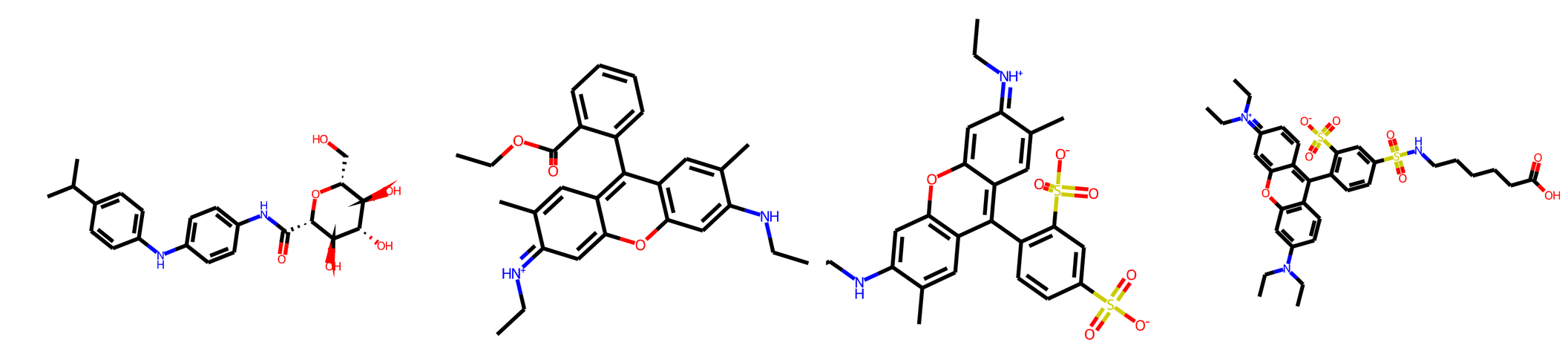
Input | RNN | Transformer | T5 | MolT5 | Ground Truth

The molecule is a member of the class of phhenylureas that is urea in which one of the nitrogens is substituted by a p-chlorophenyl group while the other is substituted by two methyl groups. It has a role as a herbicide, a xenobiotic and an environmental contaminant. It is a member of monochlorobenzenes and a member of phenylureas.

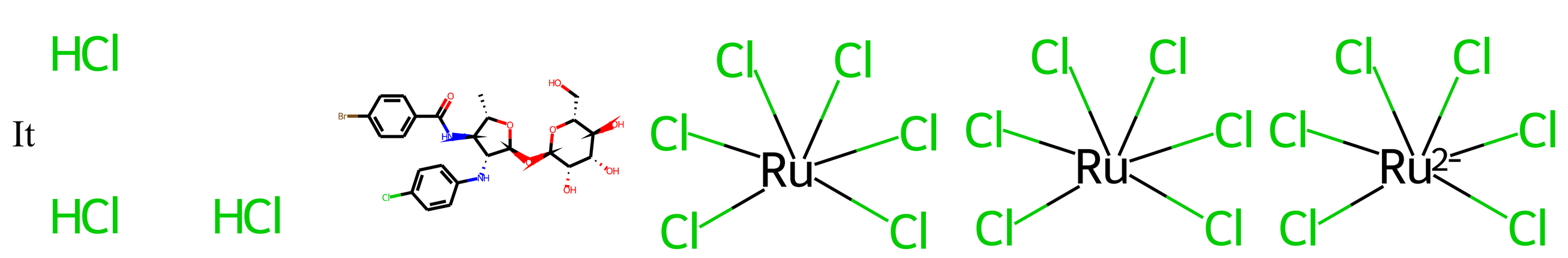


The molecule is a sulfonated xanthene dye of absorption wavelength 573 nm and emission wavelength 591 nm. It has a role as a fluorochrome.

Invalid

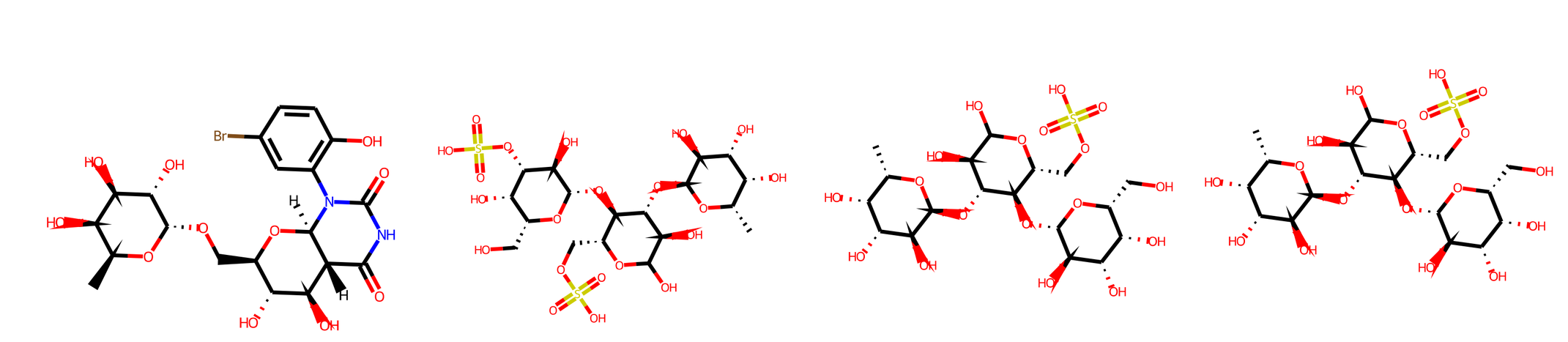


The molecule is a perchlorometallate anion having six chlorines and ruthenium(IV) as the metal component. It is a perchlorometallate anion and a ruthenium coordination entity.



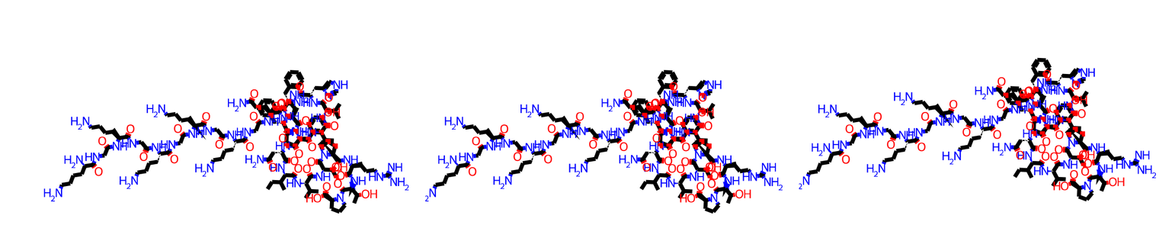
The molecule is a trisaccharide derivative that consists of 6-sulfated D-glucose having an alpha-L-fucosyl residue attached at position 3 and a beta-D-galactosyl residue attached at position 4. It has a role as an epitope. It is a trisaccharide derivative and an oligosaccharide sulfate.

Invalid

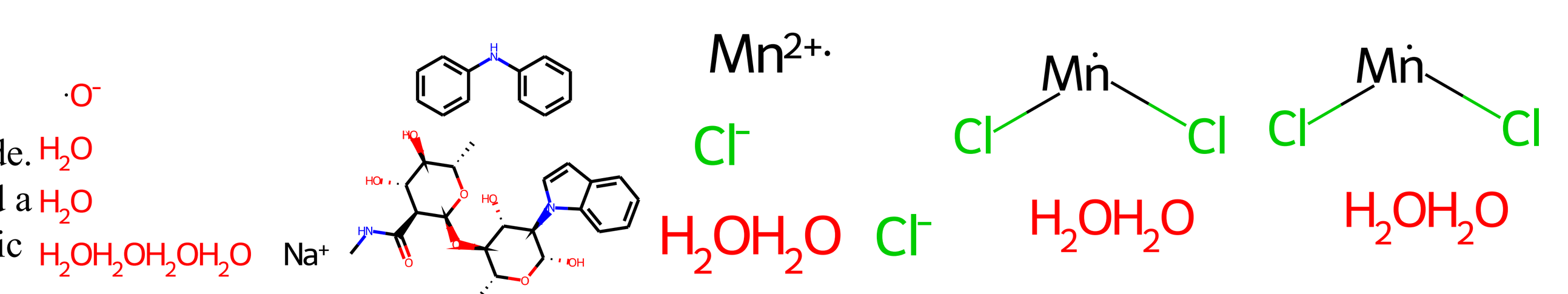


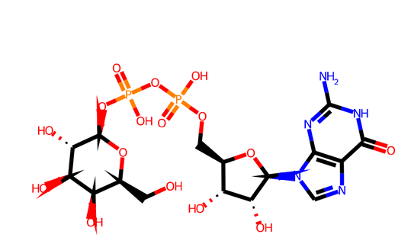
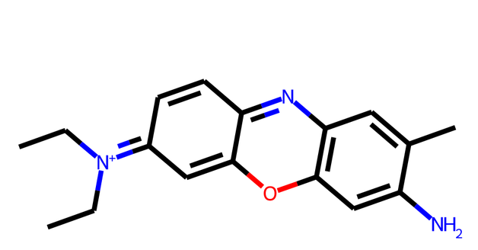
The molecule is a linear 27-membered polypeptide comprising the sequence Lys-Gly-Lys-Gly-Lys-Gly-Lys-Gly-Lys-Gly-Glu-Asn-Pro-Val-Val-His-Phe-Phe-Tyr-Asn-Ile-Val-Thr-Pro-Arg-Thr-Pro. Corresponds to the sequence of the myelin basic protein 83-99 (MBP83-99) immunodominant epitope with the lysyl residue at position 91 replaced by tyrosyl [MBP83-99(Y(91))] and with an (L-lysylglycyl)5 [(KG5)] linker attached to the glutamine(83) (E(83)) residue.

Invalid | Invalid

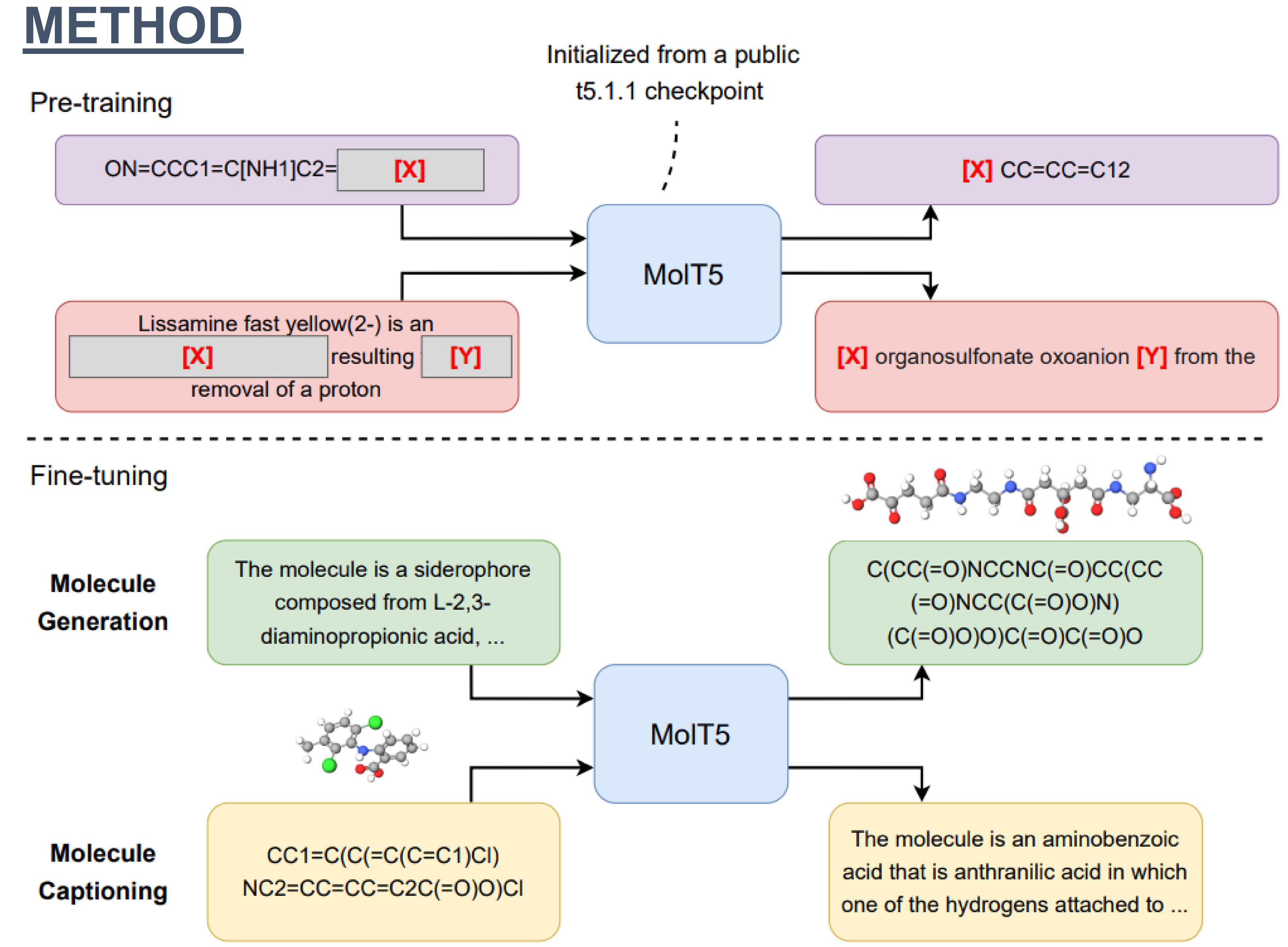


The molecule is a hydrate that is the dihydrate form of manganese(II) chloride. It has a role as a MRI contrast agent and a nutraceutical. It is a hydrate, an inorganic chloride and a manganese coordination entity.



Input	RNN	Transformer	T5	MolT5	Ground Truth
	the molecule is a gdp - d - glucoside - a - [...]	the molecule is the stable isotope of helium with relative atomic mass 3. 016029. the least abundant (0. 000137 atom percent) isotope of naturally occurring helium.	The molecule is a GDP-D-glucose in which the anomeric centre of the pyranose fragment has alpha-configuration. It is a GDP-D-glucose and a ribonucleoside 5'-diphosphate-alpha-D-glucose. It is a conjugate acid of a GDP-alpha-D-glucose(2-).	The molecule is a GDP-L-galactose in which the anomeric oxygen is on the same side of the fucose ring as the methyl substituent. It has a role as a plant metabolite and a mouse metabolite. It is a conjugate acid of a GDP-beta-L-galactose(2-).	The molecule is a GDP-L-galactose having beta-configuration at the anomeric centre of the L-galactose fragment. It is a conjugate acid of a GDP-beta-L-galactose(2-).
	the molecule is a cationic fluorescent dye having 2, 3 - dimethyl - 1, 2, 3, 4, 6 - tetrahydro - 1h - 1, 2, 3, 4, 6 - tetrahydropyridin - 1 - yl] amino } amino group, respectively. it has a role as a fluorochrome.	the molecule is a deuterated compound that is is is is is an isotopologue of chloroform in which the four hydrogen atoms have been replaced by deuterium. it is a deuterated compound and an alpha, omega - dicarboxylic acid.	The molecule is a quaternary ammonium ion and a member of phenanthridines. It has a role as an intercalator and a fluorochrome.	The molecule is an organic cation that is phenoxazin-5-ium substituted by amino and methylamino groups at positions 3 and 7 respectively. The chloride salt is the histological dye 'azure C'.	The molecule is an organic cation that is phenoxazin-5-ium substituted by methyl, amino and diethylamino groups at positions 2, 3 and 7 respectively. The tetrachlorozincate salt salt is the histological dye 'brilliant cresyl blue'.

METHOD



Results

Model	BLEU-2	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	Text2Mol
Ground Truth							0.609
RNN	0.251	0.176	0.450	0.278	0.394	0.363	0.426
Transformer	0.061	0.027	0.204	0.087	0.186	0.114	0.057
T5-Small	0.501	0.415	0.602	0.446	0.545	0.532	0.526
MolT5-Small	0.519	0.436	0.620	0.469	0.563	0.551	0.540
T5-Base	0.511	0.423	0.607	0.451	0.550	0.539	0.523
MolT5-Base	0.540	0.457	0.634	0.485	0.578	0.569	0.547
T5-Large	0.558	0.467	0.630	0.478	0.569	0.586	0.563
MolT5-Large	**0.594**	**0.508**	**0.654**	**0.510**	**0.594**	**0.614**	**0.582**

Table 1: Molecule captioning results on the test split of CheBI-20. Rouge scores are F1 values.

Model	BLEU↑	Exact↑	Levenshtein↓	MACCS FTS↑	RDK FTS↑	Morgan FTS↑	FCD↓	Text2Mol↑	Validity↑
Ground Truth	1.000	1.000	0.0	1.000	1.000	1.000	0.0	0.609	1.0
RNN	0.652	0.005	38.09	0.591	0.400	0.362	0.223	0.409	0.542
Transformer	0.499	0.000	57.66	0.480	0.320	0.217	0.379	0.277	**0.906**
T5-Small	0.741	0.064	27.703	0.704	0.578	0.525	0.213	0.479	0.608
MolT5-Small	0.755	0.079	25.988	0.703	0.568	0.517	0.198	0.482	0.721
T5-Base	0.762	0.069	24.950	0.731	0.605	0.545	0.177	0.499	0.660
MolT5-Base	0.769	0.081	24.458	0.721	0.588	0.529	0.185	0.496	0.772
T5-Large	0.854	0.279	16.721	0.823	0.731	0.670	0.117	0.552	0.902
MolT5-Large	**0.854**	**0.311**	**16.071**	**0.834**	**0.746**	**0.684**	**0.116**	**0.554**	0.905

Table 2: Molecule generation results on the test split of CheBI-20. Except for BLEU, Exact, Levenshtein, and Validity, other metrics are computed using only syntactically valid molecules, as in (Campos and Ji, 2021).

CONCLUSIONS

We propose MolT5, a framework for pretraining a model on both molecules and natural language. It enables two new tasks– molecule captioning and text-based molecule generation. These have promising applications for enabling semantic, functional-level control of molecules, democratizing access to AI technologies for designing molecules, and enabling the design of task-specific custom molecules.

ACKNOWLEDGEMENTS

This research is based upon work supported by the Molecule Maker Lab Institute: An AI Research Institutes program supported by NSF under Award No. 2019897 and NSF No. 2034562. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.