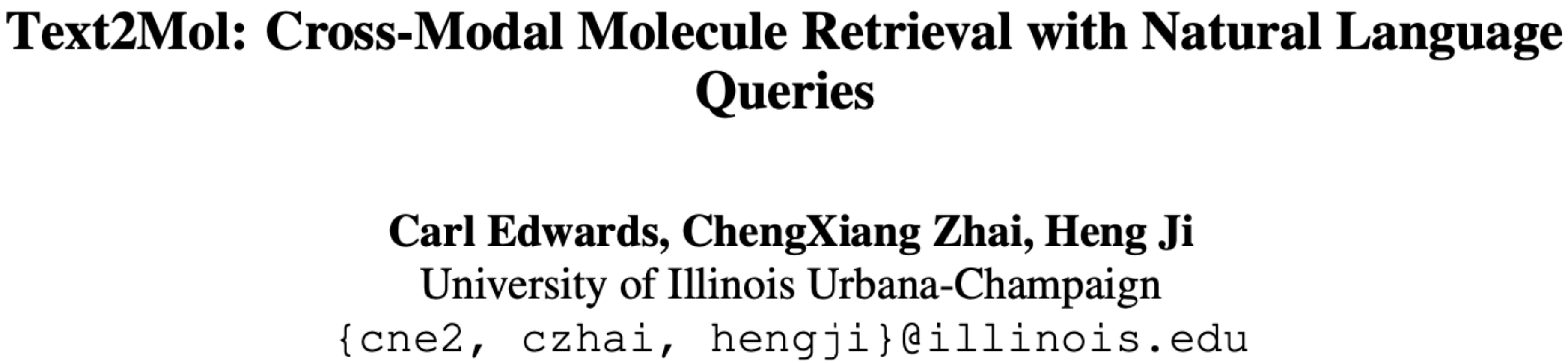
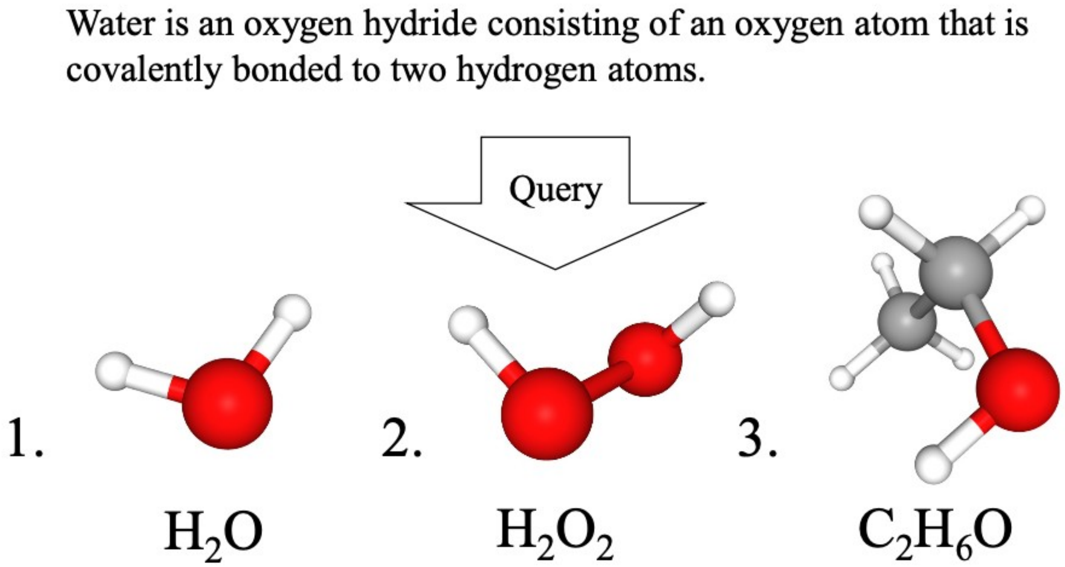


Text2Mol: Cross-Modal Molecule Retrieval with Natural Language Queries

Carl Edwards, ChengXiang Zhai, Heng Ji
University of Illinois Urbana-Champaign
{cne2, czhai, hengji}@illinois.edu

1. Goals & contributions –

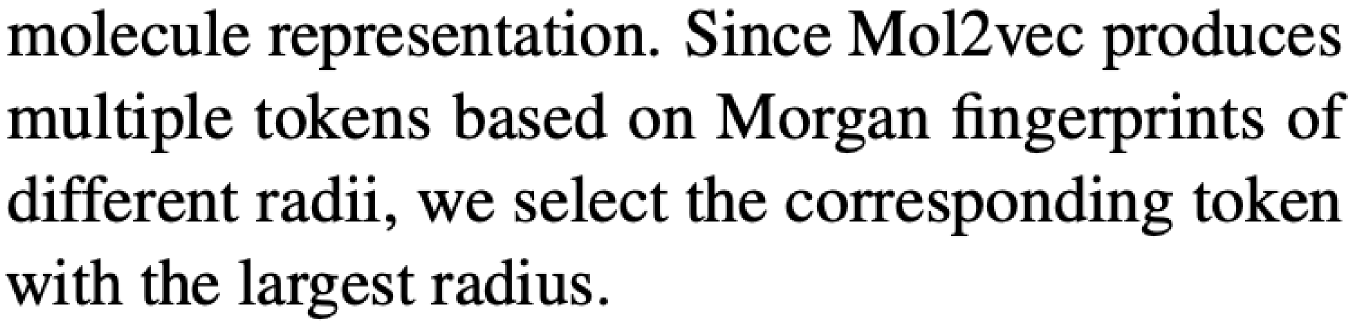
- Fast *molecule retrieval* given text prompts.
- Natural language + molecules = *different* modalities.
- Proposed method – *“a multimodal embedding approach for constructing an aligned semantic space between these two types of data to allow for cross-modal retrieval”.*
- Big + on **explainability**.
- (Contribution) *new task* **Text2Mol**, new CLIR.
- (Contribution) *explainability* through *association rules*.
- (Contribution) new dataset.



Water is an oxygen hydride consisting of an oxygen atom that is covalently bonded to two hydrogen atoms.

2. Related work –

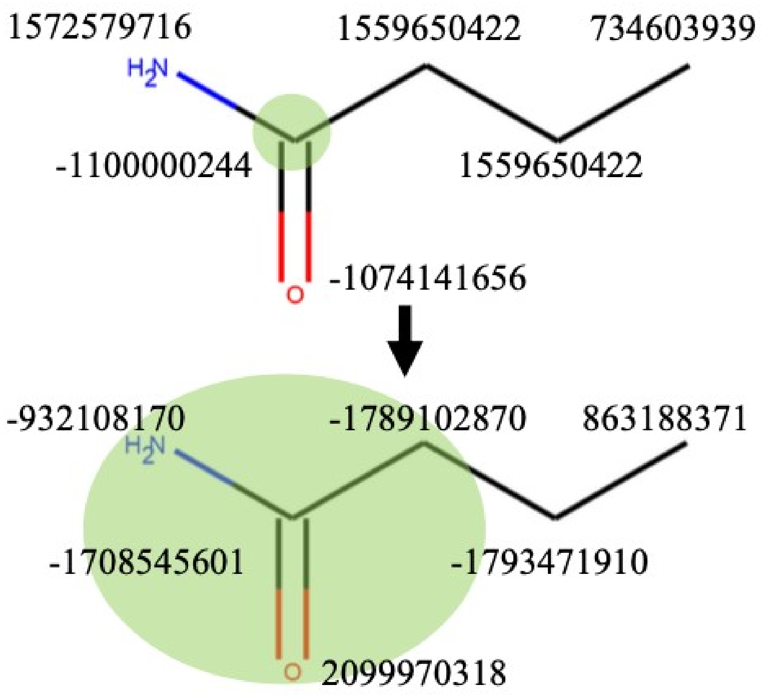
- *Multimedia Representation* (CLIP)
- *Molecule Representation*
- *Description retrieval*
- *CLIR*



molecule representation. Since Mol2vec produces multiple tokens based on Morgan fingerprints of different radii, we select the corresponding token with the largest radius.

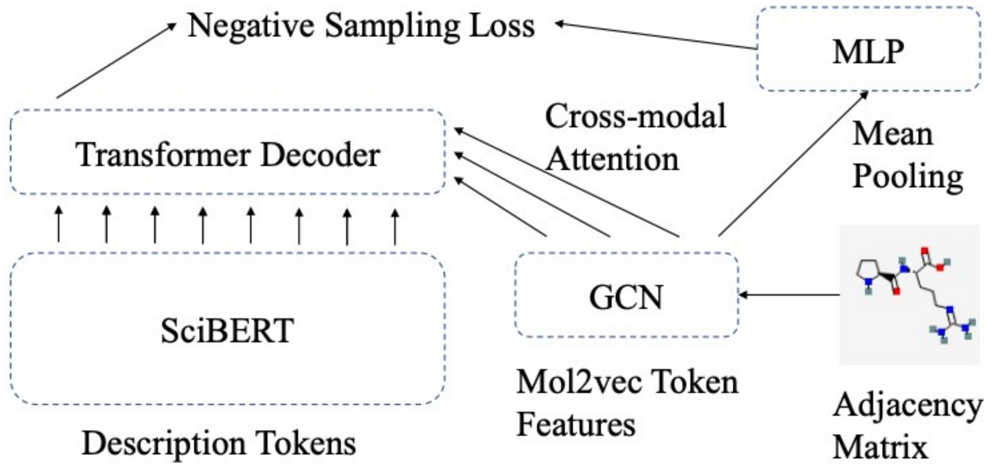
3. Method + model –

- Separate text encoder + molecule encoder + cosine-similarity to rank embeddings. (CLIP?)
- Text Encoder – *SciBERT + linear projection.*
- Molecule Encoder – (*Mol2Vec + MLP*) + (*Mol2Vec + GCN*).
- **Mol2vec.**
- For paper, *Mol2vec -> 2-layer MLP.*
- **Tough to capture large-graph info.,** so use **GCN.**
- *o/p -> 3-layer GCN -> global mean-pooling -> 2-layer MLP.*



3. Method + model –

- Cross-modality?



4. Loss – (not CLIP!)

$$L(m,t) = CCE(e^{\tau} m t^{T}, I_n) + CCE(e^{\tau} t m^{T}, I_n)$$ ❌

5. Cross-modal re-ranking –

$$supp(r) = \sum_{p \in P} \sum_{\substack{t' \in p_t \\ m' \in p_m}} \mathbb{1}_{\substack{t=t' \\ m=m'}} a_{t',m'}$$

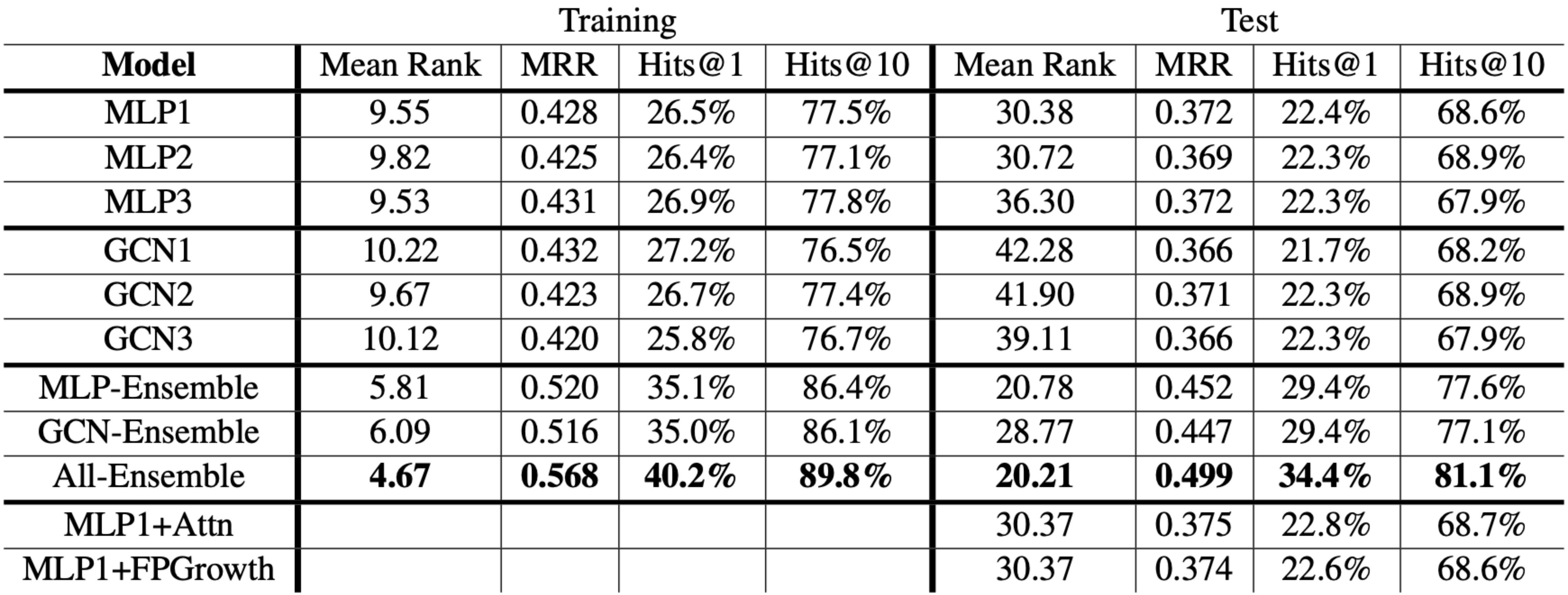
$$conf(t \implies m) = \frac{supp(t,m)}{\sum_{t' \in T} supp(t',m)}$$

$$S(a,b) = \alpha \cos(a,b) + (1-\alpha) AR(a,b)$$

6. Ensemble –

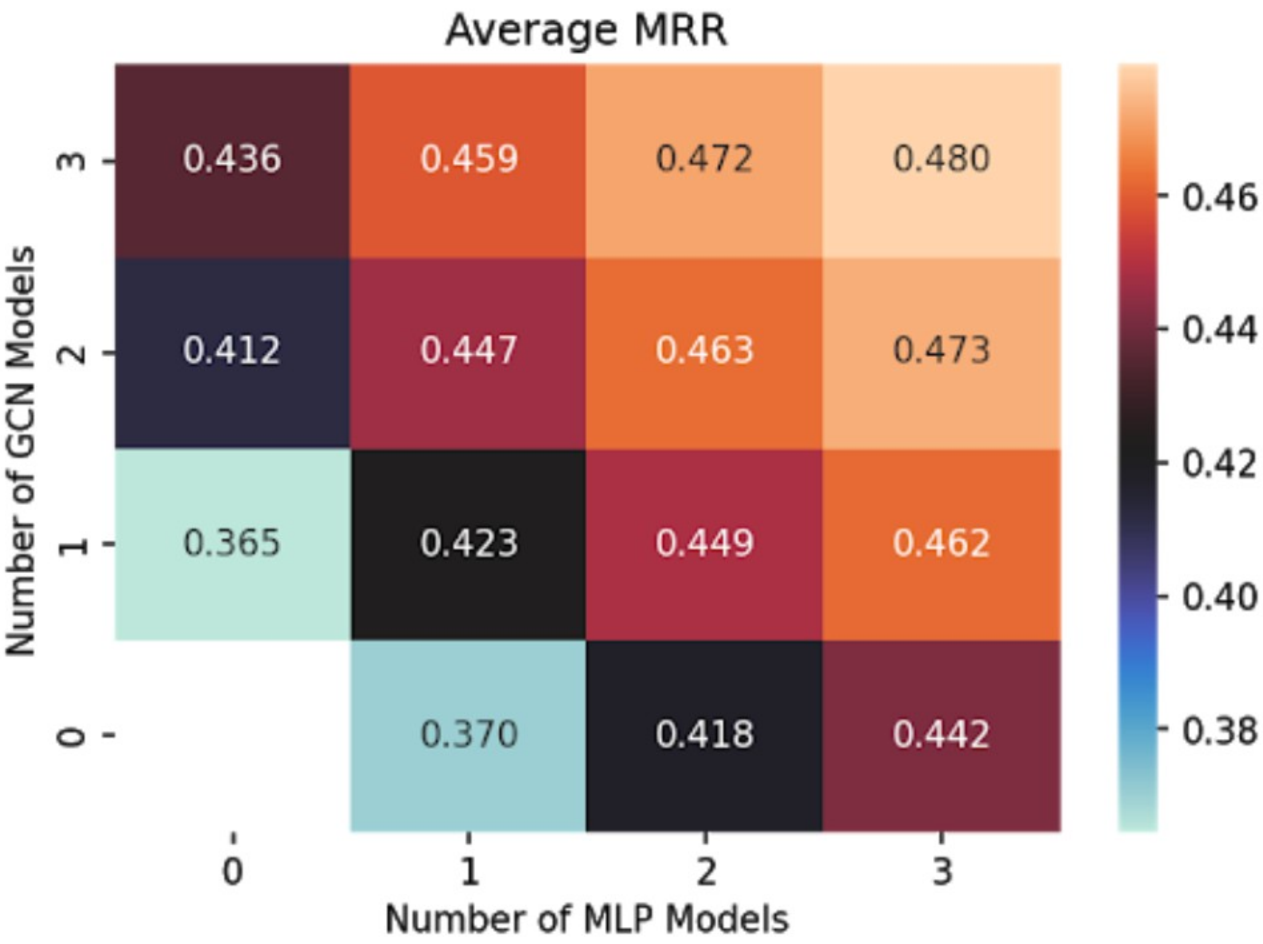
$$S(m) = \sum_i w_i R_i(m) \qquad s.t. \sum_i w_i = 1$$

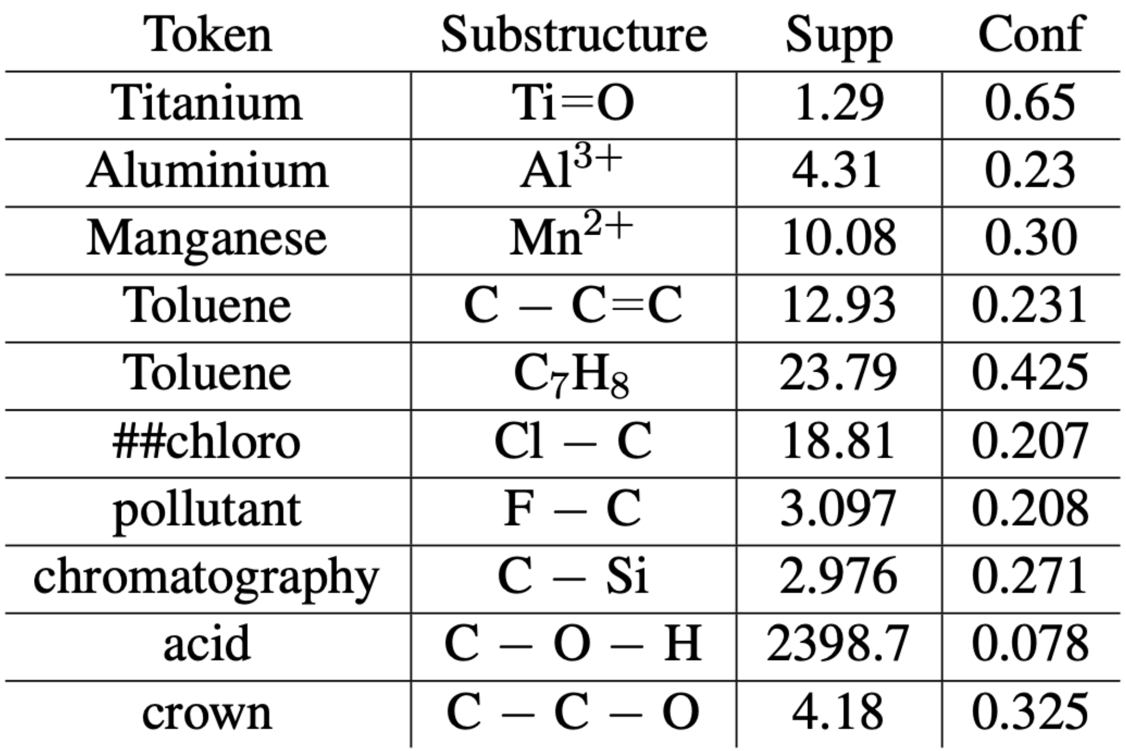
7. Experiments –



	Training				Test			
Model	Mean Rank	MRR	Hits@1	Hits@10	Mean Rank	MRR	Hits@1	Hits@10
MLP1	9.55	0.428	26.5%	77.5%	30.38	0.372	22.4%	68.6%
MLP2	9.82	0.425	26.4%	77.1%	30.72	0.369	22.3%	68.9%
MLP3	9.53	0.431	26.9%	77.8%	36.30	0.372	22.3%	67.9%
GCN1	10.22	0.432	27.2%	76.5%	42.28	0.366	21.7%	68.2%
GCN2	9.67	0.423	26.7%	77.4%	41.90	0.371	22.3%	68.9%
GCN3	10.12	0.420	25.8%	76.7%	39.11	0.366	22.3%	67.9%
MLP-Ensemble	5.81	0.520	35.1%	86.4%	20.78	0.452	29.4%	77.6%
GCN-Ensemble	6.09	0.516	35.0%	86.1%	28.77	0.447	29.4%	77.1%
All-Ensemble	**4.67**	**0.568**	**40.2%**	**89.8%**	**20.21**	**0.499**	**34.4%**	**81.1%**
MLP1+Attn					30.37	0.375	22.8%	68.7%
MLP1+FPGrowth					30.37	0.374	22.6%	68.6%

7. Experiments –

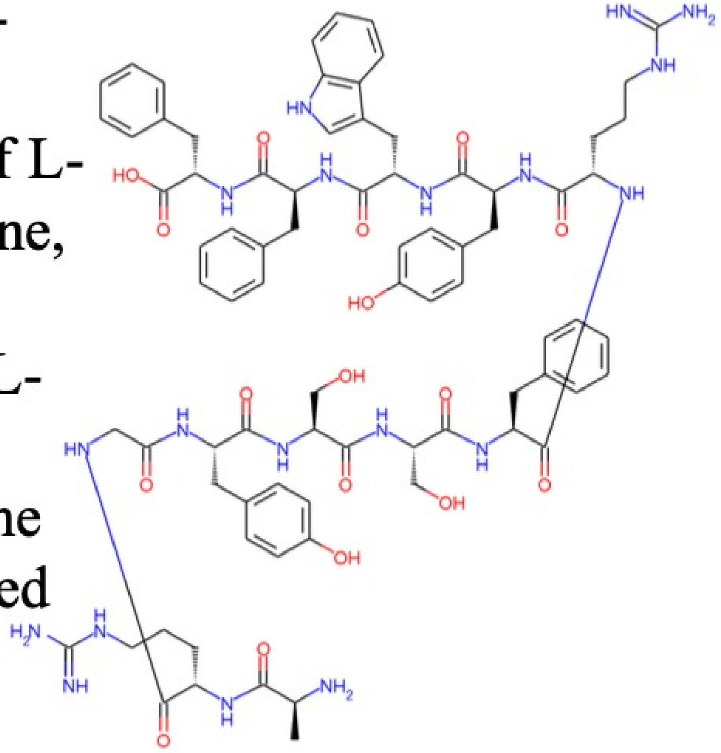


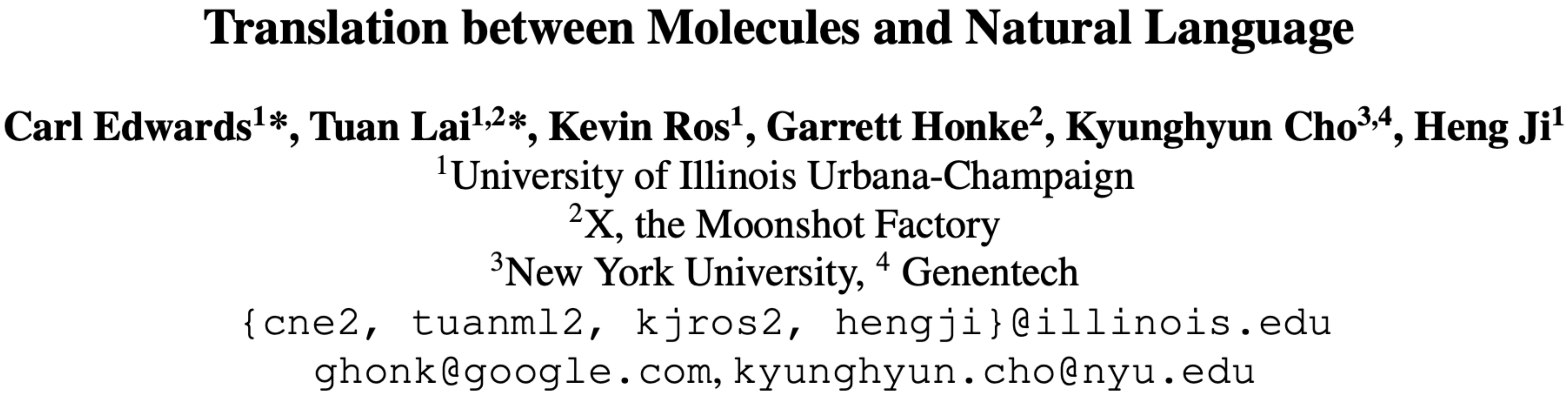


Token	Substructure	Supp	Conf
Titanium	$Ti{=}O$	1.29	0.65
Aluminium	Al^{3+}	4.31	0.23
Manganese	Mn^{2+}	10.08	0.30
Toluene	$C-C{=}C$	12.93	0.231
Toluene	C_7H_8	23.79	0.425
##chloro	$Cl-C$	18.81	0.207
pollutant	$F-C$	3.097	0.208
chromatography	$C-Si$	2.976	0.271
acid	$C-O-H$	2398.7	0.078
crown	$C-C-O$	4.18	0.325

7. Experiments –

Argyssfrywff: Ala-Arg-Gly-Tyr-Ser-Ser-Phe-Arg-Tyr-Trp-Phe-Phe is an oligopeptide composed of L-alanine, L-arginine, glycine, L-tyrosine, L-serine, L-serine, L-phenylalanine, L-arginine, L-tyrosine, L-trytophan, L-phenylalanine and L-phenylalanine joined in sequence by peptide linkages.



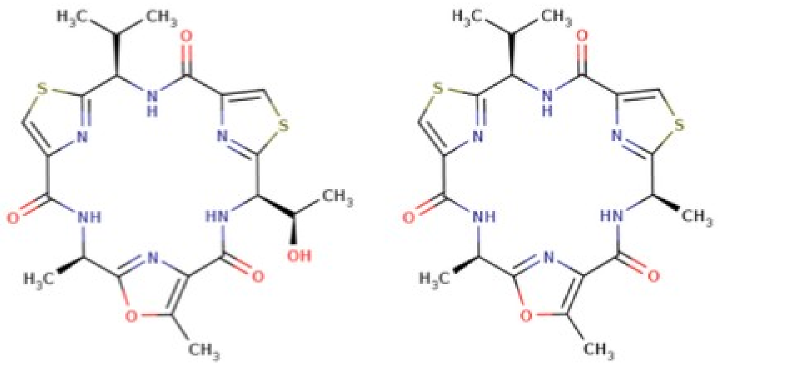


Translation between Molecules and Natural Language

Carl Edwards[1]*, Tuan Lai[1,2]*, Kevin Ros[1], Garrett Honke[2], Kyunghyun Cho[3,4], Heng Ji[1]
[1]University of Illinois Urbana-Champaign
[2]X, the Moonshot Factory
[3]New York University, [4] Genentech
{cne2, tuanml2, kjros2, hengji}@illinois.edu
ghonk@google.com, kyunghyun.cho@nyu.edu

1. Goals & contributions –

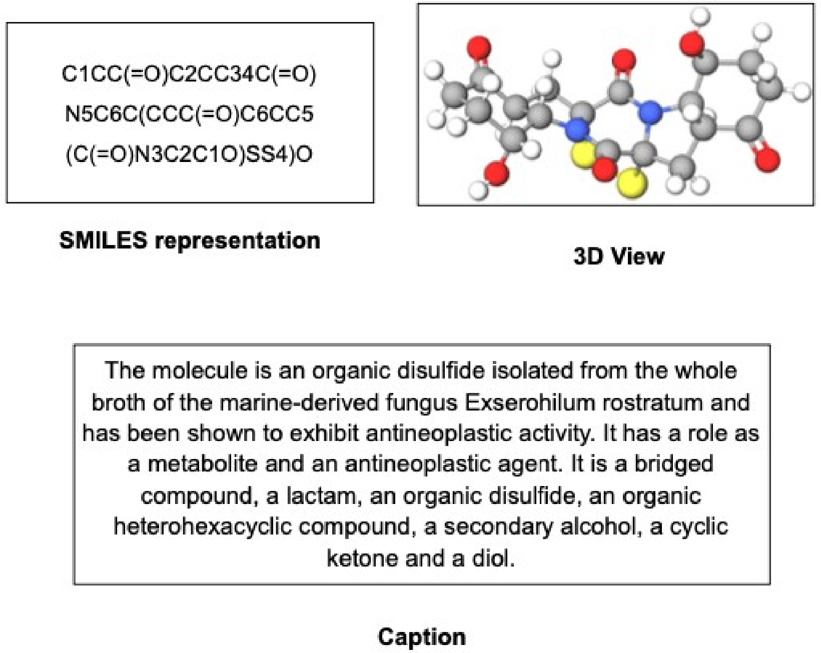
- *"we pursue an ambitious goal of translating between molecules and language by proposing two new tasks:* ***molecule captioning*** *and* ***text-guided de novo molecule generation****".*



Target

Prediction

The molecule is an eighteen-membered homodetic cyclic peptide which is isolated from Oscillatoria sp. and exhibits antimalarial activity against the W2 chloroquine-resistant strain of the malarial parasite, Plasmodium falciparum. It has a role as a metabolite and an antimalarial. It is a homodetic cyclic peptide, a member of 1,3-oxazoles, a member of 1,3-thiazoles and a macrocycle.



C1CC(=O)C2CC34C(=O)
N5C6C(CCC(=O)C6CC5
(C(=O)N3C2C1O)SS4)O

SMILES representation

3D View

The molecule is an organic disulfide isolated from the whole broth of the marine-derived fungus Exserohilum rostratum and has been shown to exhibit antineoplastic activity. It has a role as a metabolite and an antineoplastic agent. It is a bridged compound, a lactam, an organic disulfide, an organic heterohexacyclic compound, a secondary alcohol, a cyclic ketone and a diol.

Caption

- Different + (tougher) from vision-language models.
- (Contribution) **MolT5** model which is much like GPT's paradigm.
- (Contribution) new evaluation metrics (eg. **Text2Mol similarity**)

2. Tasks –

- *Molecule Captioning*
- *De novo Molecule Generation*

3. Evaluation –

$$\operatorname{lev}(a,b)=\begin{cases}|a| & \text{if } |b|=0,\\ |b| & \text{if } |a|=0,\\ \operatorname{lev}\big(\operatorname{tail}(a),\operatorname{tail}(b)\big) & \text{if } \operatorname{head}(a)=\operatorname{head}(b),\\ 1+\min\begin{cases}\operatorname{lev}\big(\operatorname{tail}(a),b\big)\\ \operatorname{lev}\big(a,\operatorname{tail}(b)\big)\\ \operatorname{lev}\big(\operatorname{tail}(a),\operatorname{tail}(b)\big)\end{cases} & \text{otherwise}\end{cases}$$

- Text2Mol Metric.
- BLEU, METEOR and ROGUE for molecule captioning evaluation.
- Novelty and Scaffold similarity ❌

  BUT instead

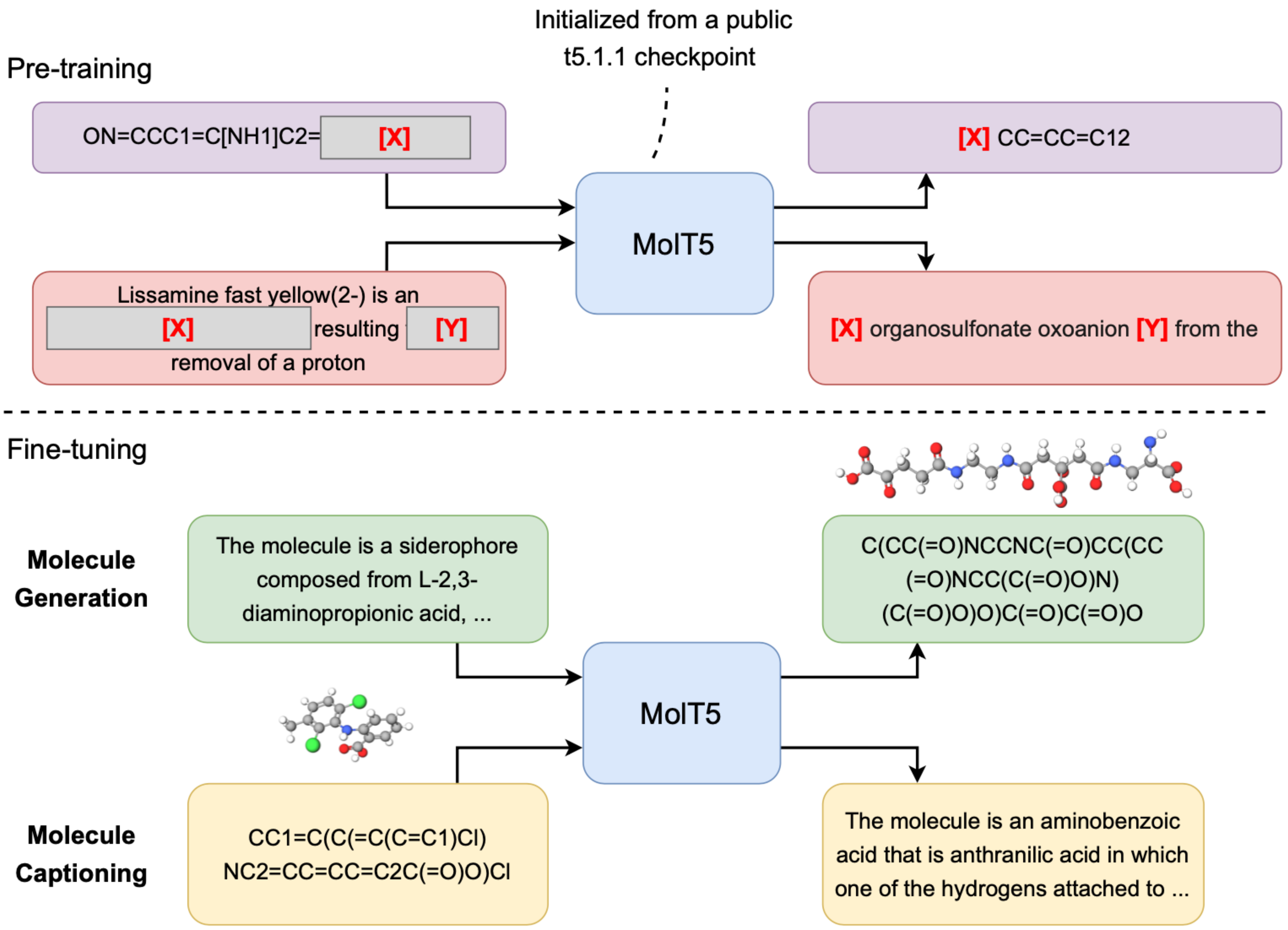
  MACCS FTS, **RDK FTS**, and **Morgan FTS**

  Levenshtein distance, **exact SMILES match** and **SMILES BLEU**

  Fréchet ChemNet Distance (FCD) $d^2((\boldsymbol{m},\boldsymbol{C}),(\boldsymbol{m}_w,\boldsymbol{C}_w))=\|\boldsymbol{m}-\boldsymbol{m}_w\|_2^2+\operatorname{Tr}\big(\boldsymbol{C}+\boldsymbol{C}_w-2(\boldsymbol{C}\boldsymbol{C}_w)^{1/2}\big).$

  Syntactic validity using RDKIT

4. MolT5 –



Initialized from a public t5.1.1 checkpoint

Pre-training

ON=CCC1=C[NH1]C2= [X]

[X] CC=CC=C12

MolT5

Lissamine fast yellow(2-) is an [X] resulting [Y] removal of a proton

[X] organosulfonate oxoanion [Y] from the

Fine-tuning

Molecule Generation

The molecule is a siderophore composed from L-2,3-diaminopropionic acid, ...

C(CC(=O)NCCNC(=O)CC(CC(=O)NCC(C(=O)O)N)(C(=O)O)O)C(=O)C(=O)O

MolT5

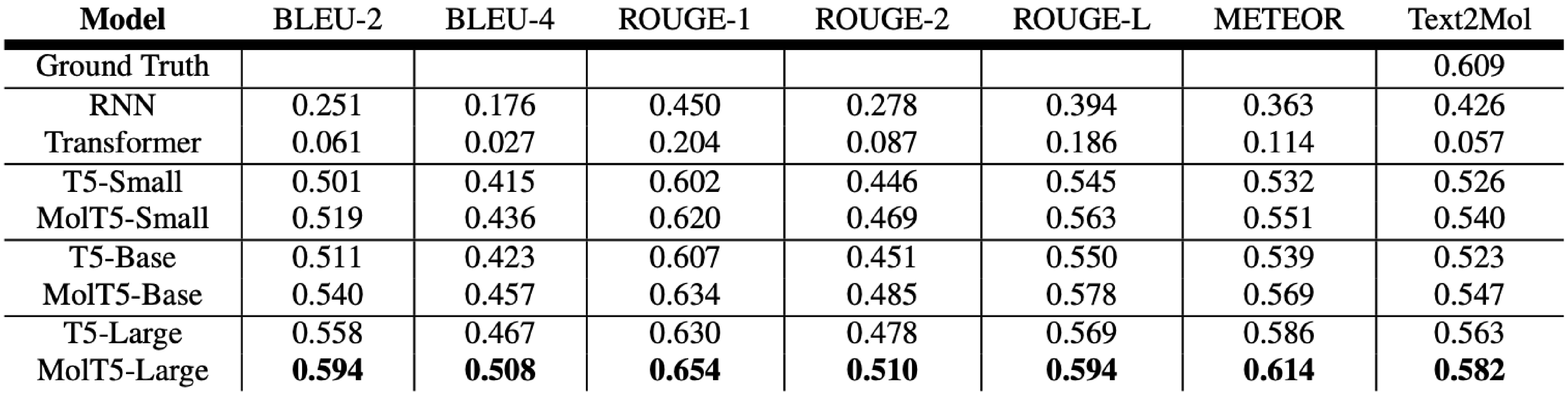
Molecule Captioning

CC1=C(C(=C(C=C1)Cl)NC2=CC=CC=C2C(=O)O)Cl

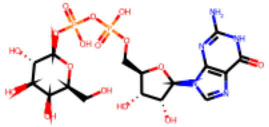
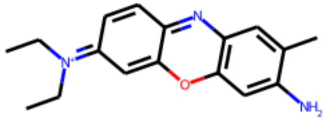
The molecule is an aminobenzoic acid that is anthranilic acid in which one of the hydrogens attached to ...

5. Experiments –

- Pre-training Data : C4 for text, ZINC-15 for molecules
- Fine-tuning Data : ChEBI-20
- Baselines : GRU, Transformer, T5
- *Molecule Captioning*

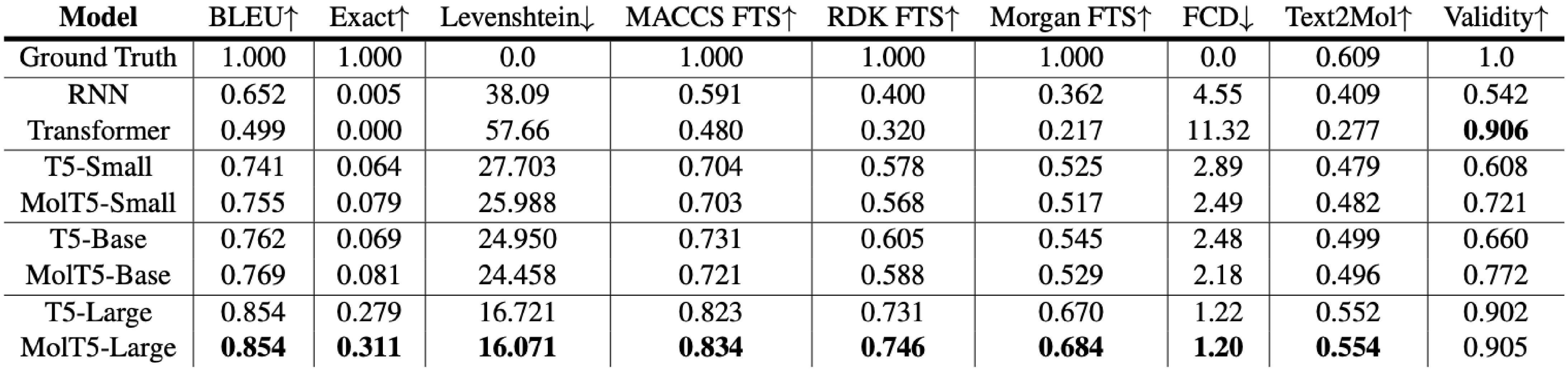


Model	BLEU-2	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	Text2Mol
Ground Truth							0.609
RNN	0.251	0.176	0.450	0.278	0.394	0.363	0.426
Transformer	0.061	0.027	0.204	0.087	0.186	0.114	0.057
T5-Small	0.501	0.415	0.602	0.446	0.545	0.532	0.526
MolT5-Small	0.519	0.436	0.620	0.469	0.563	0.551	0.540
T5-Base	0.511	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**

	Input	RNN	Transformer	T5	MolT5	Ground Truth
1		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-).
2	^{39}Ar	the molecule is stable metallic [...]	the molecule is the stable isotope of thallium with relative atomic mass 202. 9723. the least abundant (29. 524 atom percent) isotope of naturally occurring thallium.	The molecule is the radioactive isotope of chromium with relative atomic mass 39.98286 and half-life of 138.376 days; the only naturally occurring isotope of chromium.	The molecule is the stable isotope of rubidium with relative atomic mass 44.955910, 100 atom percent natural abundance and nuclear spin 7/2.	The molecule is a trace radioisotope of argon with atomic mass of 38.964313 and a half-life of 269 years. It has a role as an isotopic tracer.
3		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 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'.

5. Experiments –

- *Text based De Novo generation*

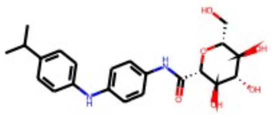
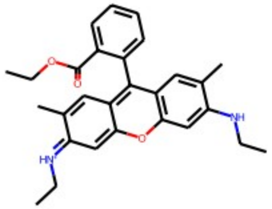
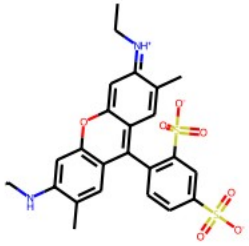
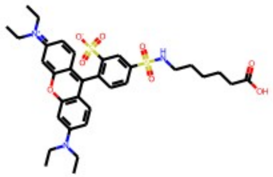
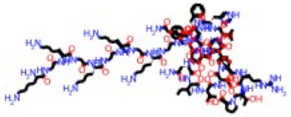
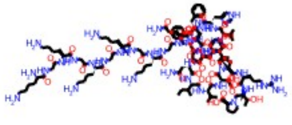
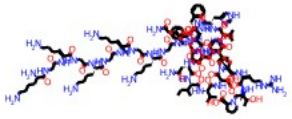
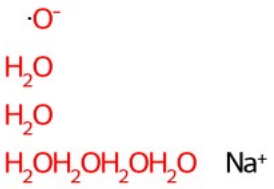
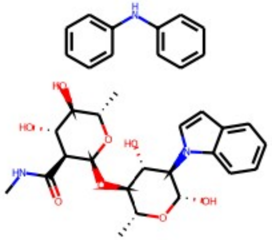
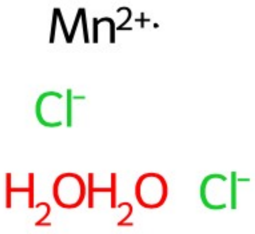
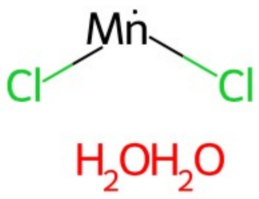
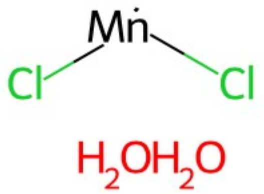


Model	BLEU↑	Exact↑	Levenshtein↓	MACCS FTS↑	RDK FTS↑	Morgan FTS↑	FCD↓	Text2Mol↑	Validity↑
Ground Truth	1.000	1.000	0.0	1.000	1.000	1.000	0.0	0.609	1.0
RNN	0.652	0.005	38.09	0.591	0.400	0.362	4.55	0.409	0.542
Transformer	0.499	0.000	57.66	0.480	0.320	0.217	11.32	0.277	**0.906**
T5-Small	0.741	0.064	27.703	0.704	0.578	0.525	2.89	0.479	0.608
MolT5-Small	0.755	0.079	25.988	0.703	0.568	0.517	2.49	0.482	0.721
T5-Base	0.762	0.069	24.950	0.731	0.605	0.545	2.48	0.499	0.660
MolT5-Base	0.769	0.081	24.458	0.721	0.588	0.529	2.18	0.496	0.772
T5-Large	0.854	0.279	16.721	0.823	0.731	0.670	1.22	0.552	0.902
MolT5-Large	**0.854**	**0.311**	**16.071**	**0.834**	**0.746**	**0.684**	**1.20**	**0.554**	0.905

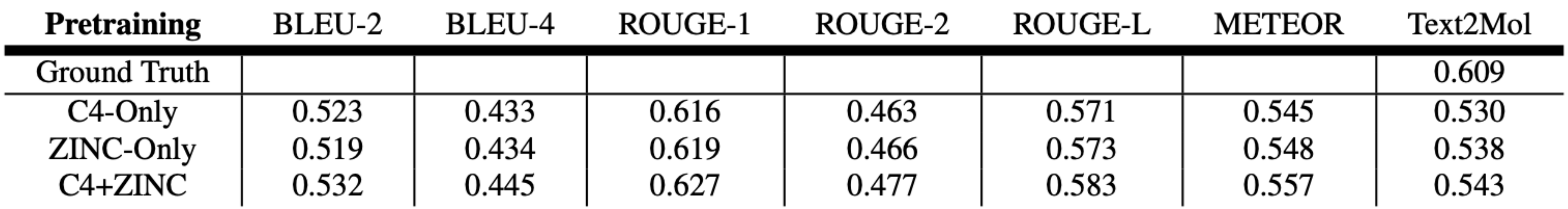
- *With generation* ***there are a LOT of properties that can be investigated based on different molecules.***

5. Experiments –

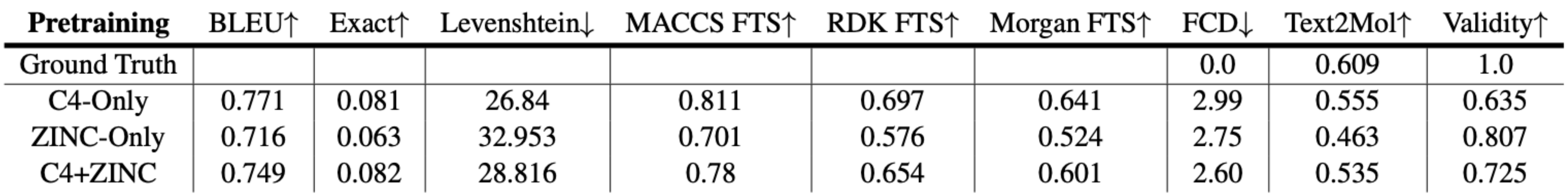
- *Text based De Novo generation*

	Input	RNN	Transformer	T5	MolT5	Ground Truth
1	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				
2	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			
3	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.					

6. Ablations –



Pretraining	BLEU-2	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	Text2Mol
Ground Truth							0.609
C4-Only	0.523	0.433	0.616	0.463	0.571	0.545	0.530
ZINC-Only	0.519	0.434	0.619	0.466	0.573	0.548	0.538
C4+ZINC	0.532	0.445	0.627	0.477	0.583	0.557	0.543



Pretraining	BLEU↑	Exact↑	Levenshtein↓	MACCS FTS↑	RDK FTS↑	Morgan FTS↑	FCD↓	Text2Mol↑	Validity↑
Ground Truth							0.0	0.609	1.0
C4-Only	0.771	0.081	26.84	0.811	0.697	0.641	2.99	0.555	0.635
ZINC-Only	0.716	0.063	32.953	0.701	0.576	0.524	2.75	0.463	0.807
C4+ZINC	0.749	0.082	28.816	0.78	0.654	0.601	2.60	0.535	0.725

My Thoughts (based off of the 10 papers we have read) –

1. **First**, we MUST involve a Knowledge Graph of some sort, basically I think all text-molecule papers till now have random out-of-context text descriptions of a molecule valid in a particular field, not general enough to generate good molecules, which is why I think they have these results. Knowledge Graph makes concrete the information LLM is supposed to learn.

2. **Second**, this paper also had statistical tests and we should do similar tests.

3. **Third**, our approach should focus on multiple tasks **and** be modular so that we can use it downstream independent of other modalities, e.g., GraphMVP doesn't need 3D info during inference time.

4. **Fourth**, and this is based off of a vision-language model I have seen, basically it would be very ideal to have 2 phases, one for representation learning and one for generative tasks, that would be very modular as well.

5. **Fifth**, for representation learning we need to come up with a good **modality interaction** module.