Text2Mol: Cross-modal Molecule Retrieval with Natural Language Queries

Carl Edwards, ChengXiang Zhai, and Heng Ji
Department of Computer Science, College of Engineering, University of Illinois at Urbana-Champaign

MOLECULE MAKER LAB INSTITUTE

NSF

INTRODUCTION

- There are hundreds of millions of molecules, so finding the right molecule for a problem can be challenging.
- Many current information retrieval systems rely only on textual descriptions of molecules, but there are more molecules than can possibly be tested in a lab and then described.
- To address this issue, it is critical to retrieve molecules directly from natural language descriptions.

TASK DEFINITION

- Given a text query and list of molecules without any reference textual information, retrieve the molecule corresponding to the query.
- We assume there is only one correct (relevant) molecule for each description, so we consider two measures for this task: Hits@1 and mean reciprocal rank (MRR).

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

Query

1. H_2O 2. H_2O_2 3. C_2H_6O

DATA

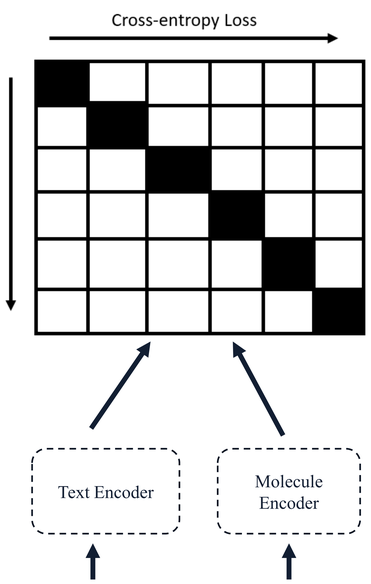
- We collect ChEBI annotations of compounds scraped from PubChem, which consists of 102,980 compound-description pairs
- Using this data, we create a dataset consisting of 33,010 pairs, which we call **ChEBI-20**, that contains descriptions of more than 20 words.
- The dataset is split into 80%/10%/10% train/validation/test splits. Models are trained on the training data, and queries are performed on all molecules in the entire dataset.

METHOD

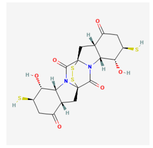
- Molecules and natural language are very different though: how can we combine them?
- In this work, we learned an aligned embedding space between molecules and natural language using symmetric contrastive loss.

Contrastive Symmetric Loss

Outer product of molecule and text batch compared to identity matrix:

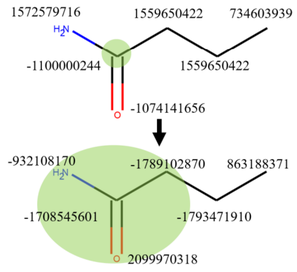


Rostratin D is an organic disulfide isolated from the whole broth of the marine-derived fungus Exserohilum rostratum and has been shown to exhibit antineoplastic activity. […] It is a bridged compound, a cyclic ketone, a lactam, an organic disulfide, an organic heterohexacyclic compound, a secondary alcohol, a dithiol and a diol.



Encoding

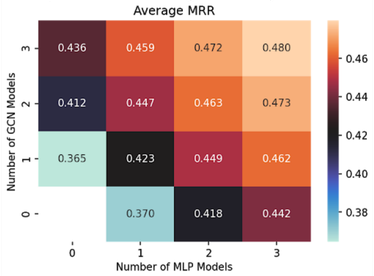
- Descriptions are encoded using SciBERT [1].
- Molecules are encoded with two methods: an MLP and GCN. The MLP builds off Mol2vec [2].
- We introduce a GCN to better incorporate the graph structure of the molecule.



- Mol2vec uses Morgan fingerprints of atoms to turn molecules into "sentences" for the Word2vec algorithm

Ensemble

- The correct molecule was very frequently ranked highly by all models.
- Incorrect molecules being ranked highly rarely occurs in multiple models (even with the same architecture).
- Averaging ranks improves performance significantly.

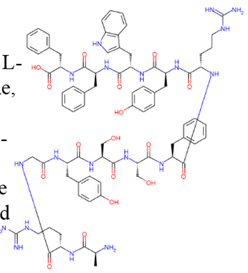


	Training				Test			
Model	Mean Rank	MRR	Hits@1	Hits@10	Mean Rank	MRR	Hits@1	Hits@10
MLP1	9.55	0.428	26.5%	77.5%	30.38	0.372	22.4%	68.6%
MLP2	9.82	0.425	26.4%	77.1%	30.72	0.369	22.3%	68.9%
MLP3	9.53	0.431	26.9%	77.8%	36.30	0.372	22.3%	67.9%
GCN1	10.22	0.432	27.2%	76.5%	42.28	0.366	21.7%	68.2%
GCN2	9.67	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%

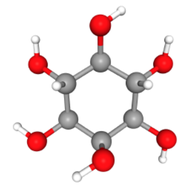
RESULTS

Retrieved Correctly:

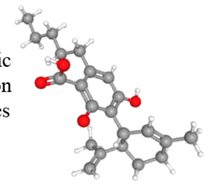
Argyssfrywff: Ala-Arg-Gly-Tyr-Ser-Ser-Phe-Arg-Tyr-Trp-Phe-Phe is an oligopeptide composed of L-alanine, L-arginine, glycine, L-tyrosine, L-serine, L-serine, L-phenylalanine, L-arginine, L-tyrosine, L-trytophan, L-phenylalanine and L-phenylalanine joined in sequence by peptide linkages.



Inositol: Myo-inositol is an inositol having myo-configuration. It has a role as a member of compatible osmolytes, a nutrient, an EC 3.1.4.11 (phosphoinositide phospholipase C) inhibitor, a human metabolite, a Daphnia magna metabolite, […]

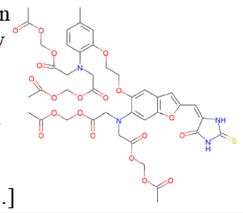


Cannabidiolate is a dihydroxybenzoate that is the conjugate base of cannabidiolic acid, obtained by deprotonation of the carboxy group. It derives from an olivetolate. It is a conjugate base of a cannabidiolic acid.



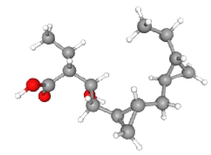
Retrieved Incorrectly:

Fura red is a 1-benzofuran substituted at position 2 by a (5-oxo-2-thioxoimidazolidin-4-ylidene)methyl group, and at C-5 and C-6 by heavily substituted oxygen and nitrogen functionalities […]

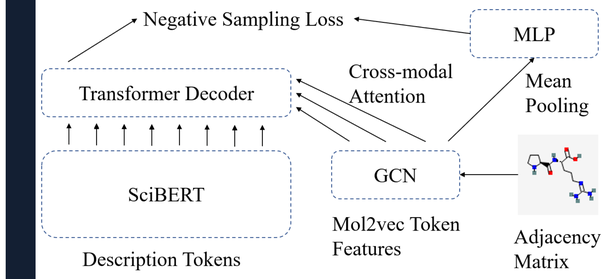


Clondronate(2-) is the dianion resulting from the removal of two protons from clondronic acid. It is a conjugate base of a clodronic acid.

An alpha-mycolic acid is a class of mycolic acids characterized by the presence of two cis cyclopropyl groups in the meromycolic chain. It is an organic molecular entity and a mycolic acid. […]



ATTENTION-BASED ASSOCIATION RULES



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

CONCLUSIONS

We show that it is possible to align molecules and their descriptions for cross-modal retrieval. We argue that natural language and molecules, while very different, are complementary sources of information that can and should be integrated together.

References

[1] Beltagy, Iz, Kyle Lo, and Arman Cohan. "SciBERT: A Pretrained Language Model for Scientific Text." Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019.

[2] Jaeger, Sabrina, Simone Fulle, and Samo Turk. "Mol2vec: unsupervised machine learning approach with chemical intuition." Journal of chemical information and modeling 58.1 (2018): 27-35.

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.

ILLINOIS