Text2Mol: Cross-modal Molecular 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

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.

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.

Encoding

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

ATTENTION BASED ASSOCIATION RULES

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


ACKNOWLEDGEMENTS

This work was supported by the Molecule Maker Lab Institute: An AI Research Institutes program supported by NSF under Award No. 2019897. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect those of the National Science Foundation.

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.

TASK DEFINITION

• Given a text query and list of molecules with out 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).

RESULTS


Retrieved Incorrectly: Furo red is a 1-benzofurfin substituted at position 2 by a (5-ctto-2-thiocinimidazolin-4-ylidenemethyl group. at C-5 and C-6 by heavily substituted oxygen and nitrogen.

Clindronate(2-) is the dianion resulting from the removal of two cations from clindamycin acid. It is a conjugate base of a clindamic acid.

An alpha-mycolic acid is a class of mycolic acids characterized by the presence of two cis cyclic glycol groups in the mycolic acid chain. It is an organic molecular entity and a mycolic acid.

ATTENTION-BASED ASSOCIATION RULES

ATTENTION-BASED ASSOCIATION RULES

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.

Encoding

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

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


ACKNOWLEDGEMENTS

This work was supported by the Molecule Maker Lab Institute: An AI Research Institutes program supported by NSF under Award No. 2019897. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect those of the National Science Foundation.