Vision Language Model is NOT All You Need: Augmentation Strategies for Molecule Language Model

Namkyeong Lee, Siddhartha Laghuvarapu,
Chanyoung Park, Jimeng Sun

KAIST DSAIL Data Science & Artificial Intelligence UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN CIKM 2024 BOISE OCTOBER 21-25

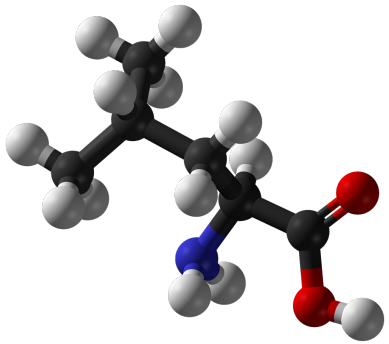
Vision Language Model is NOT All You Need: Augmentation Strategies for Molecule Language Model

Namkyeong Lee, Siddhartha Laghuvarapu,
Chanyoung Park, Jimeng Sun

KAIST DSAIL Data Science & Artificial Intelligence UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN CIKM 2024 BOISE OCTOBER 21-25

TABLE OF CONTENTS

BACKGROUND MOLECULE LANGUAGE MODELS



Molecule

Human language Description

Molecule Language Models

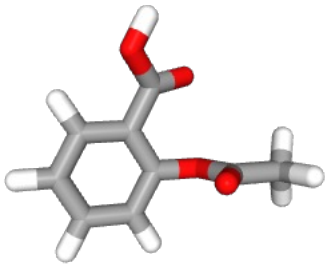
Understanding molecules in human language format

Major Research Direction

How to align molecules and their description better?

→ Mostly inspired by the recent VLM papers

BACKGROUND MOLECULE LANGUAGE MODELS



Aspirin

Aspirin is a commonly used drug for the treatment of pain and fever

Penicillin G Sodium is the sodium salt form of benzylpenicillin

CLIP-Style Pre-training

- Molecule and its corresponding description as a positive pair
- Descriptions of other molecules as negative pairs

Aligning the pairs in representation space

Molecule and its description to be close in representation space

MOTIVATION UNIQUE CHALLENGES IN MOLECULE LANGUAGE MODEL

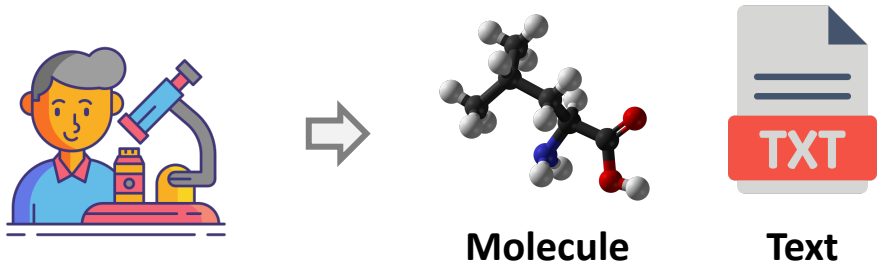
Challenge 1. Limited Quantity of Molecule-Text Pair

Vision Language Model (VLM)

"Scale is Everything" view point

→ Massive amounts of crawled image-text pairs

→ Correlation between number of pairs and model performance



Molecule Language Model (MoLM)

Expensive wet-lab experiments are required for generating molecule-text pairs

How to increase the number of molecule-text pairs without expensive wet-lab experiments?

MOTIVATION UNIQUE CHALLENGES IN MOLECULE LANGUAGE MODEL

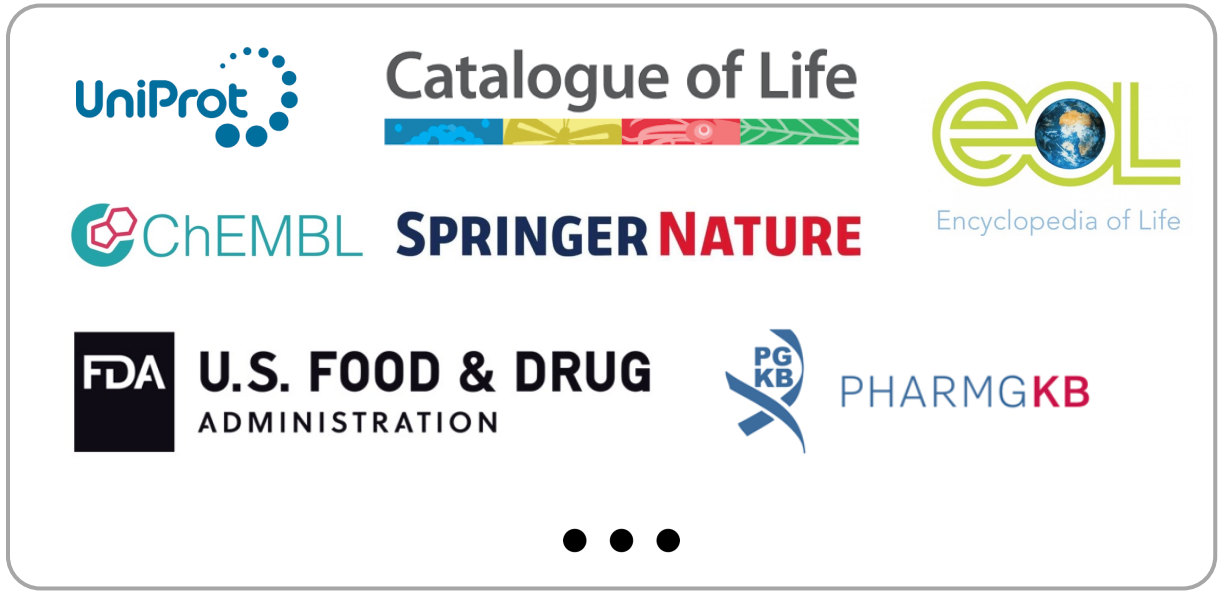
Challenge 2. Different Expertise with Diverse Interests

PubChem

Largest molecule database

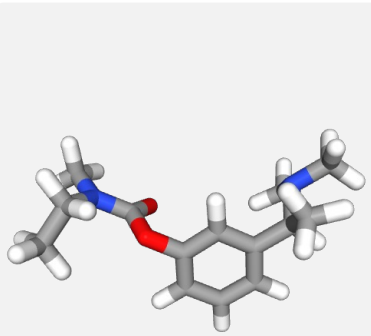
Over 94M molecules

From 1014 different data sources



MOTIVATION UNIQUE CHALLENGES IN MOLECULE LANGUAGE MODEL

Challenge 2. Different Expertise with Diverse Interests



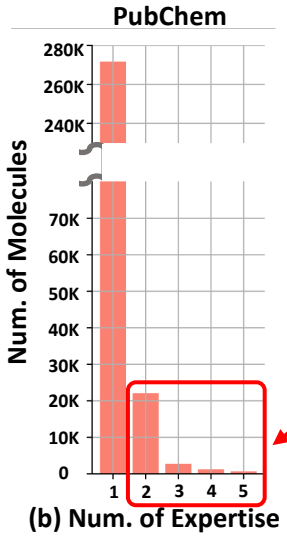
(a) Rivastigmine

FDA Pharm Class

This molecule is a Cholinesterase Inhibitor. The mechanism of action of rivastigmine is as a Cholinesterase Inhibitor.

LiverTox

This molecule is an oral acetylcholinesterase inhibitor used for therapy of Alzheimer disease.



(b) Num. of Expertise

Specialized areas of experts

→ Typically restrict their knowledge to selective group of molecules

→ Numerous molecules have **missing expertise** across various experts

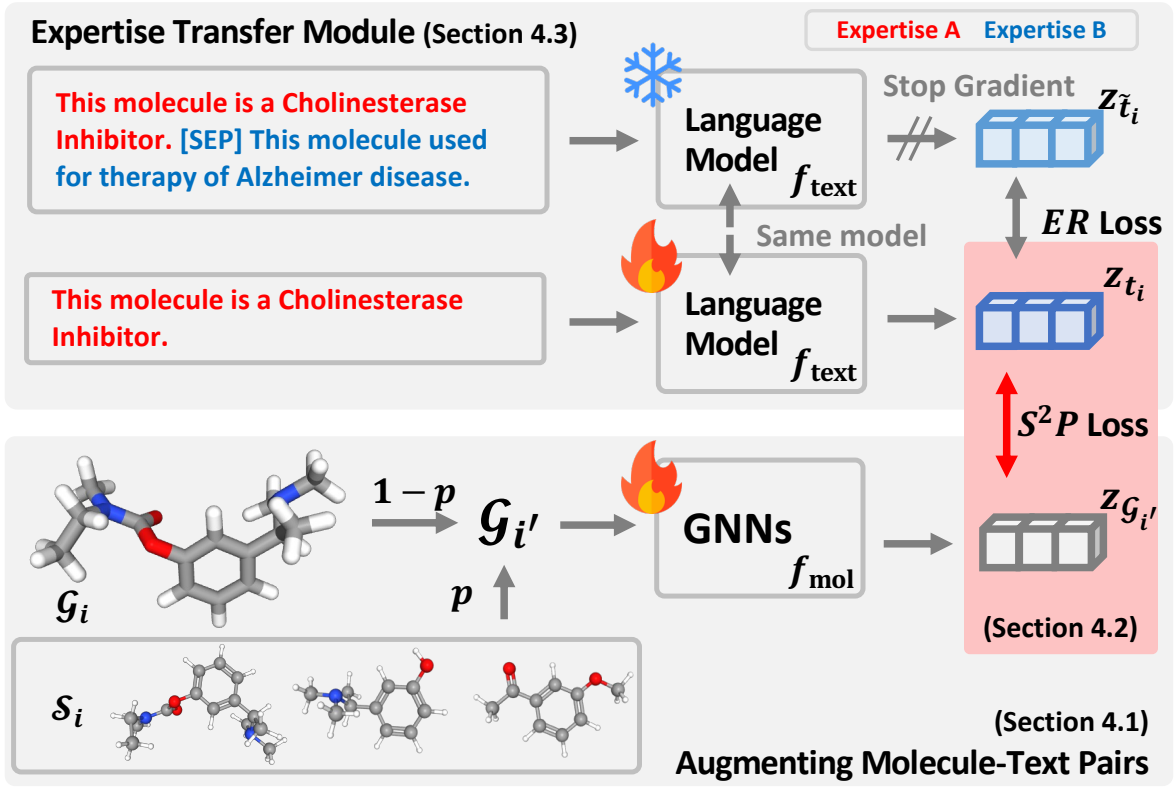
Only 27K Molecules out of 299K molecules have descriptions from multiple experts

How to fill the missing expertise based on existing data?

Vision Language Model is NOT All You Need: Augmentation Strategies for Molecule Language Model

AMOLE: Augmenting molecule-text pair and transferring expertise between the molecules

AMOLE AUGMENTING MOLECULE-TEXT PAIR AND TRANSFERRING EXPERTISE BETWEEN THE MOLECULES



Overall model architecture

Augmenting Molecule-Text Pairs

How to increase the number of molecule-text pairs?

Structural Similarity Preserving Loss

How to align augmented pairs more precisely?

Expertise Transfer Module

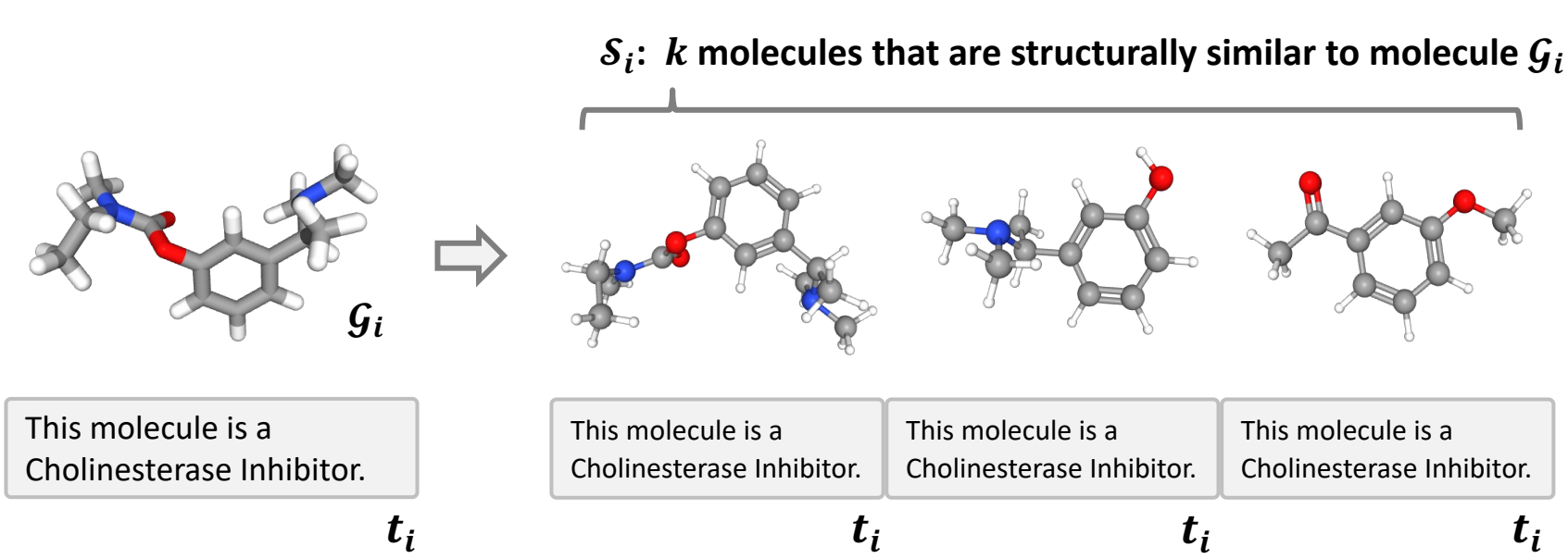
How to transfer the expertise between molecules?

AMOLE AUGMENTING MOLECULE-TEXT PAIR AND TRANSFERRING EXPERTISE BETWEEN THE MOLECULES

Augmenting Molecule-Text Pairs

Share textual descriptions among molecules with structural resemblances

→ ”Structurally similar molecules tend to display analogous biological activities”



Tanimoto Similarity

$$s_{ij} = \frac{|\mathrm{fp}_i \cap \mathrm{fp}_j|}{|\mathrm{fp}_i| + |\mathrm{fp}_j| - |\mathrm{fp}_i \cap \mathrm{fp}_j|}$$

FP: Finger Print

AMOLE AUGMENTING MOLECULE-TEXT PAIR AND TRANSFERRING EXPERTISE BETWEEN THE MOLECULES

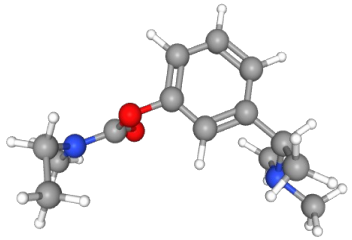
Structural Similarity Preserving Loss

Shared description t_i is not specifically written for the molecules in $\boldsymbol{S_i}$

→ Preserve structural similarity between $\boldsymbol{G_i}$ and $\boldsymbol{G_{j'}}$ in molecule-text joint space

→ The Tanimoto similarity between $\boldsymbol{G_i}$ and $\boldsymbol{G_{j'}}$ as a pseudo label for contrastive learning

Original Contrastive Learning



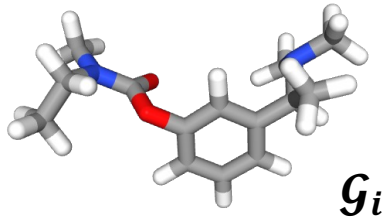
$\boldsymbol{G_{j'}}$

This molecule is a Cholinesterase Inhibitor.

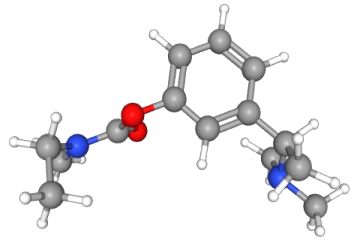
$\boldsymbol{t_i}$

Label: 1 (Positive Pair)
Label: 0 (Negative Pair)

Structural Similarity Preserving



$\boldsymbol{G_i}$



$\boldsymbol{G_{j'}}$

This molecule is a Cholinesterase Inhibitor.

$\boldsymbol{t_i}$

Label: $y_{ij'}^{t\to m} = \dfrac{\exp(s_{ij'}/\tau_1)}{\sum_{k'=1}^{N_{\text{batch}}} \exp(s_{ik'}/\tau_1)}$ s_{ij} : Tanimoto Similarity

AMOLE AUGMENTING MOLECULE-TEXT PAIR AND TRANSFERRING EXPERTISE BETWEEN THE MOLECULES

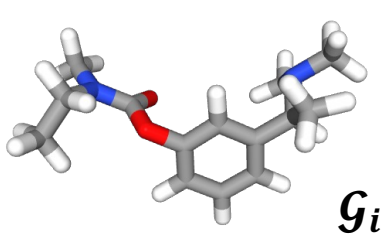
Structural Similarity Preserving Loss

Shared description t_i is not specifically written for the molecules in $\boldsymbol{S_i}$

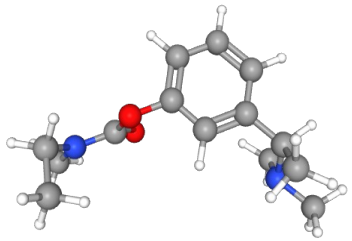
→ Preserve structural similarity between $\boldsymbol{G_i}$ and $\boldsymbol{G_{j'}}$ in molecule-text joint space

→ The Tanimoto similarity between $\boldsymbol{G_i}$ and $\boldsymbol{G_{j'}}$ as a pseudo label for contrastive learning

Structural Similarity Preserving



$\boldsymbol{G_i}$



$\boldsymbol{G_{j'}}$

This molecule is a Cholinesterase Inhibitor.

$\boldsymbol{t_i}$

Label: $y_{ij'}^{t\to m} = \frac{\exp(s_{ij'}/\tau_1)}{\sum_{k'=1}^{N_{\text{batch}}} \exp(s_{ik'}/\tau_1)}$ s_{ij} : Tanimoto Similarity

Similarity in representation space

$$\hat{y}_{ij'}^{t\to m} = \frac{\exp(\mathtt{sim}(z_{t_i}, z_{\mathcal{G}_{j'}})/\tau_2)}{\sum_{k'=1}^{N_{\text{batch}}} \exp(\mathtt{sim}(z_{t_i}, z_{\mathcal{G}_{k'}})/\tau_2)}$$

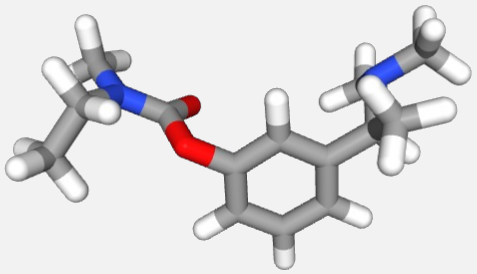
Structural Similarity Preserving (S2P) Loss

$$\mathcal{L}_{S^2P}^{t\to m} = -\frac{1}{N_{Batch}} \sum_{i=1}^{N_{batch}} \sum_{j'=1}^{N_{batch}} y_{ij'}^{t\to m} \log \hat{y}_{ij'}^{t\to m}$$

AMOLE AUGMENTING MOLECULE-TEXT PAIR AND TRANSFERRING EXPERTISE BETWEEN THE MOLECULES

Expertise Transfer Module

Transferring the expertise gained from molecules with extensive description to those with less description



(a) Rivastigmine

FDA Pharm Class

This molecule is a Cholinesterase Inhibitor. The mechanism of action of rivastigmine is as a Cholinesterase Inhibitor.

LiverTox

This molecule is an oral acetylcholinesterase inhibitor used for therapy of Alzheimer disease.

"Different areas of expertise are interrelated"

Cholinesterase inhibitor

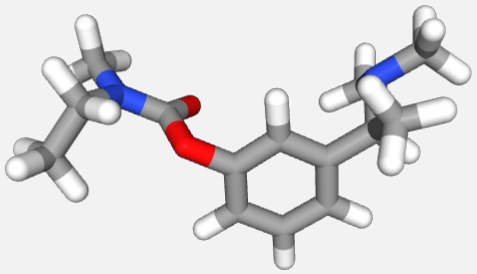
Improve communication between nerve cells
by increasing levels of Acetylcholine in the nervous system
→ Can be therapy of Alzheimer disease

FDA Pharm Classes ⇨ **LiverTox**

AMOLE AUGMENTING MOLECULE-TEXT PAIR AND TRANSFERRING EXPERTISE BETWEEN THE MOLECULES

Expertise Transfer Module

Transferring the expertise gained from molecules with extensive description to those with less description



(a) Rivastigmine

FDA Pharm Class t_i

This molecule is a Cholinesterase Inhibitor. The mechanism of action of rivastigmine is as a Cholinesterase Inhibitor.

LiverTox $t_{i'}$

This molecule is an oral acetylcholinesterase inhibitor used for therapy of Alzheimer disease.

Expertise Reconstruction (ER) Loss

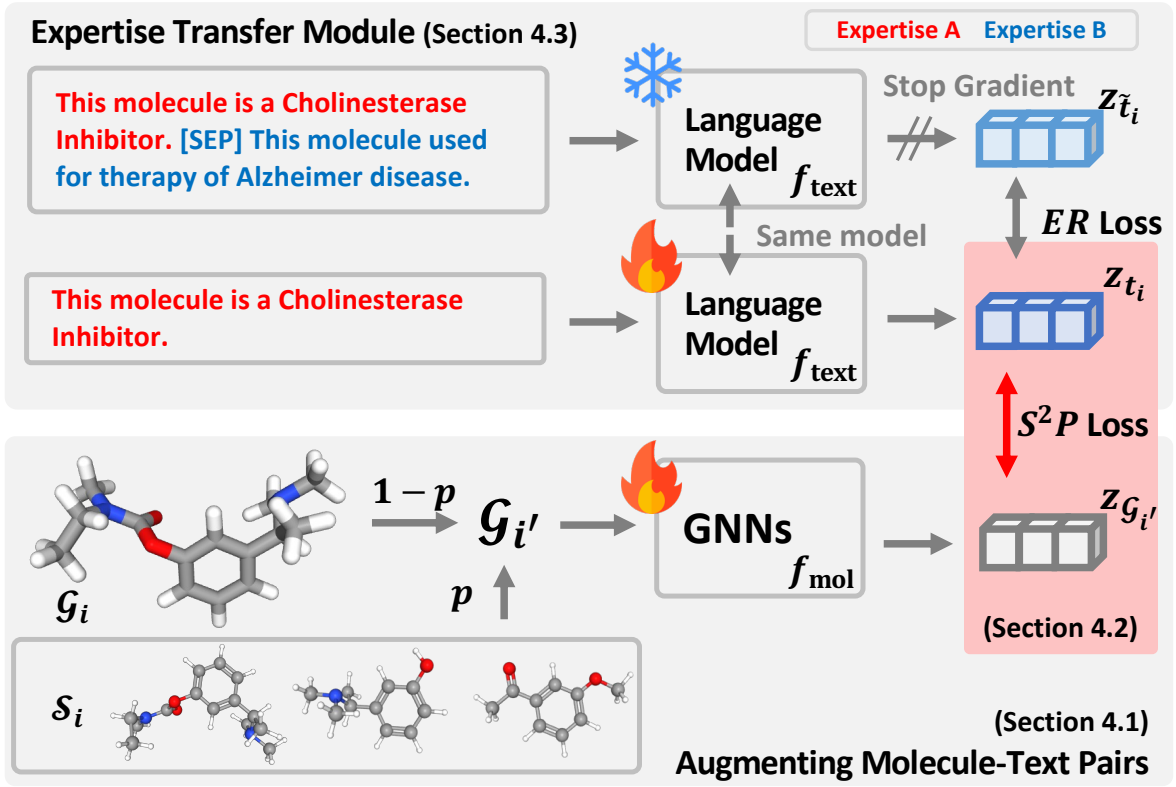
Reconstruction target $\tilde{t}_i$ by concatenating two descriptions

$\tilde{t}_i = t_i$ **[SEP]** $t_{i'}$

$$\mathcal{L}_{ER} = -\frac{1}{N_{Batch}} \sum_{i=1}^{N_{batch}} \| f_{\text{text}}(t_i) - \text{SG}(f_{\text{text}}(\tilde{t}_i)) \|_2^2$$

Model behaves as if there exists extensive expertise available, enabling reliable predictions even with lack of detailed description

AMOLE AUGMENTING MOLECULE-TEXT PAIR AND TRANSFERRING EXPERTISE BETWEEN THE MOLECULES



Overall model architecture

Final Model Training Loss

$$\mathcal{L} = \mathcal{L}_{S^2P} + \alpha \cdot \mathcal{L}_{ER}$$

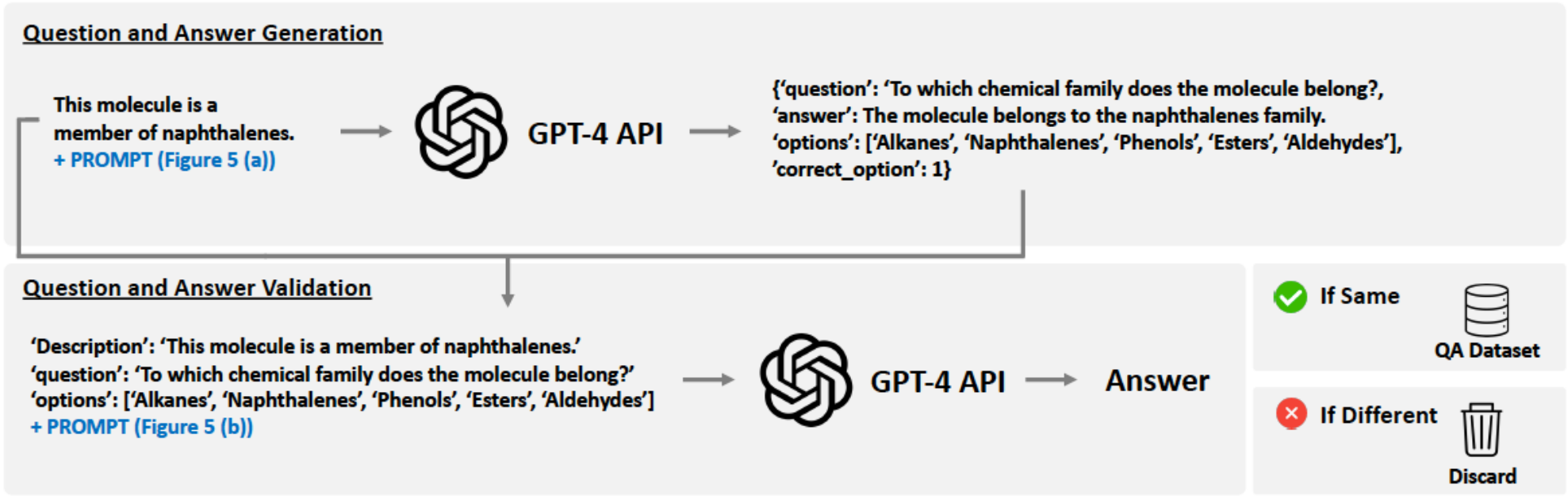
EXPERIMENTS ZERO-SHOT QUESTION AND ANSWERING

Task description

Given molecule and question, choose the correct answer from the options

Dataset Generation

Instruct GPT-4 to generate multiple-choice question comprising five options, all derived from the given textual description



EXPERIMENTS ZERO-SHOT QUESTION AND ANSWERING

Task description

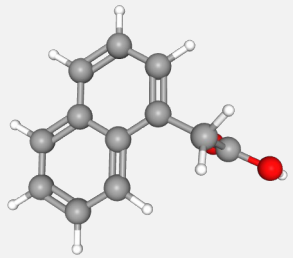
Given molecule and question, choose the correct answer from the options

Original Description
This molecule is a member of naphthalenes

GPT-Generated Questions
To which Chemical Family does the molecule belong?

GPT-Generated Options
[Alkanes, Naphtalenes, Phenols, Esters, Aldehydes]

Formulate QA Task as Retrieval Task **"Harder version of Retrieval Tasks"**



To which chemical family does the molecule belong? Alkanes
To which chemical family does the molecule belong? Naphtalenes
To which chemical family does the molecule belong? Phenols
To which chemical family does the molecule belong? Esters
To which chemical family does the molecule belong? Aldehydes

	SMILES				Graph		
	MolT5	BioT5	KV-PLM	Molecule STM	MoMu	Molecule STM	AMOLE
Descr.	24.84	27.54	30.07	36.11	38.31	37.97	**39.26**
Pharma.	22.49	27.03	26.68	29.60	29.85	29.52	**31.58**

Observations

AMOLE consistently outperforms baseline models
→ Extract more intricate information from the slight variations
→ Attribute to its ability in inferencing related expertise!

EXPERIMENTS ZERO-SHOT VIRTUAL SCREENING

Virtual Screening

Computational technique to search large libraries of compounds quickly
to identify the compounds most likely to have desired properties

Task description

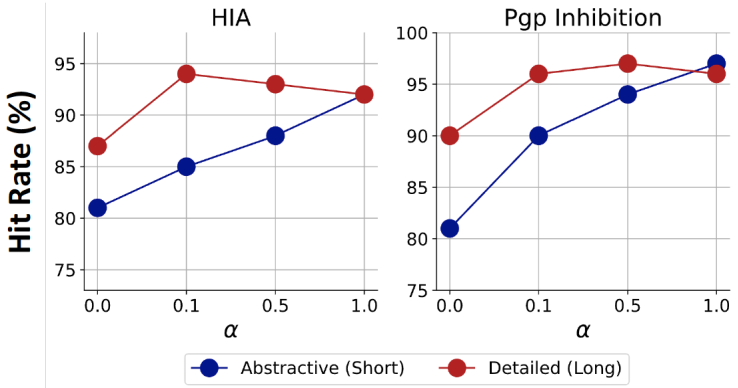
Model's proficiency in virtual screening drugs by supplying a textual description of a desired property
Two distinct descriptions for each property: Abstractive and Detailed

Dataset	Textual Description
HIA	Human intestinal absorption (HIA)
	The molecule is positive w.r.t. a property that is defined as 'the ability of the body to be absorbed from the human gastrointestinal system into the bloodstream of the human body'
Pgp Inhibition	P-glycoprotein Inhibition
	This molecule is known to inhibit P-glycoprotein, which is an ABC transporter protein involved in intestinal absorption, drug metabolism, and brain penetration, and its inhibition can seriously alter a drug's bioavailability and safety.
DILI	Inducing liver injury
	This molecule induces liver injury that is most commonly caused by Amoxicillin clavulanate isoniazid, and nonsteroidal anti-inflammatory drugs.
VDR	Vitamin D receptor
	This molecule is active w.r.t. Vitamin D receptor. The best pharmacophore hypothesis contains one hydrogen bond acceptor (A), one hydrogen bond donor (D) and two hydrophobic regions (H).

Procedure

1. Obtain the representation of textual description of the desired property
2. Top 100 molecules that are closest to the textual representation
3. Calculate the hit ratio (%)

EXPERIMENTS ZERO-SHOT VIRTUAL SCREENING

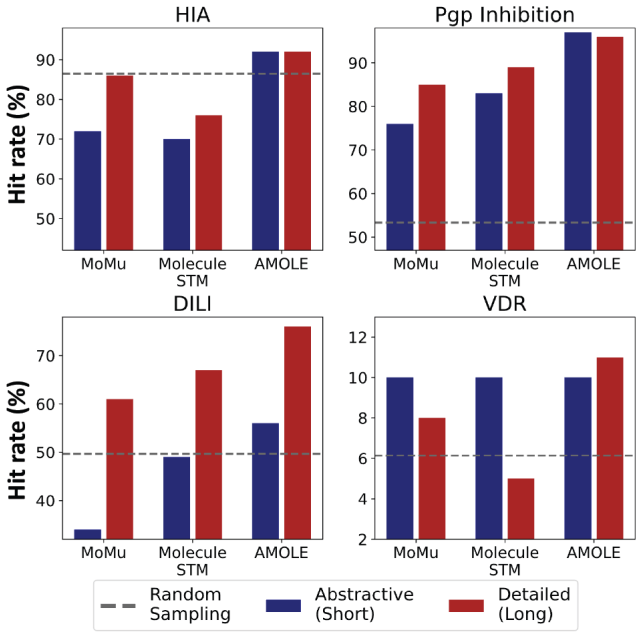


Virtual Screening

Computational technique to search large libraries of compounds quickly to identify the compounds most likely to have desired properties

Task description

Model's proficiency in virtual screening drugs by supplying a textual description of a desired property
Two distinct descriptions for each property: Abstractive and Detailed



Observations

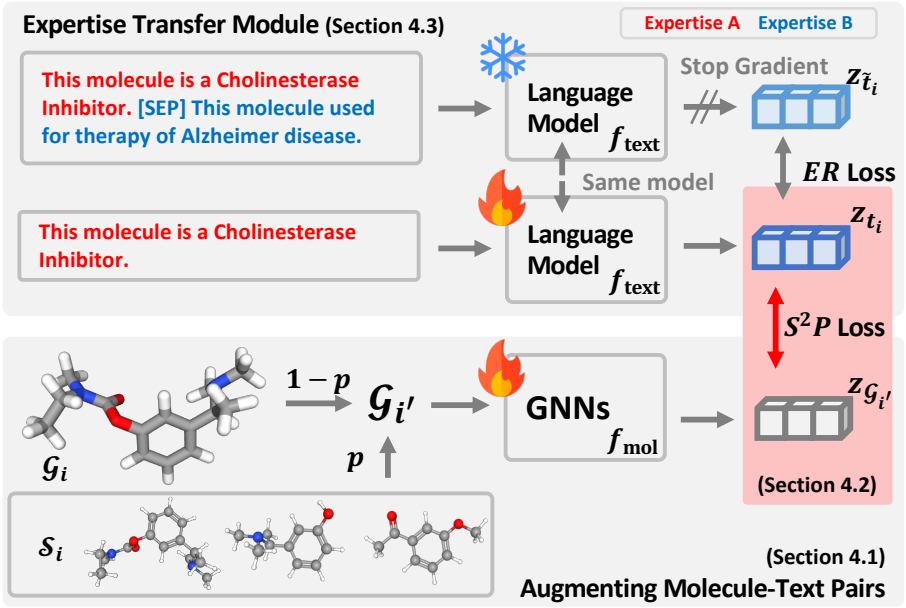
AMOLE shows notably robust performance compared to the baseline methods
→ Regardless of whether the description is abstractive or detailed
→ Attribute to Expertise Transfer module, which equips the model with the ability to deduce related information even when only a abstract level of detail is provided

Sensitivity Analysis

Detailed Prompt: Consistent Model Performance
Abstractive Prompt: Significantly varies according to the choice of α
→ As α increases, AMOLE's effectiveness improves
→ Expertise transfer module effectively allows the model to infer related information

CONCLUSION

In this paper, we propose training strategies for molecule language model



Augmenting Molecule-Text Pairs
Share the textual descriptions among molecules with structural resemblances

Structural Similarity Preserving Loss
Preserve molecules' structural similarity in molecule-text joint space

Expertise Transfer Module
Transferring the expertise gained from molecules with extensive descriptions to those with less description

Extensive experiments
- demonstrate the efficacy of AMOLE in grasping the nuances of molecules and their textual descriptions
- Highlight promising applicability for practical use in the drug discovery process

THANK YOU!

[Full Paper] https://arxiv.org/pdf/2407.09043

[Source Code] https://github.com/Namkyeong/AMOLE

[Author Email] namkyeong96@kaist.ac.kr

Paper Link

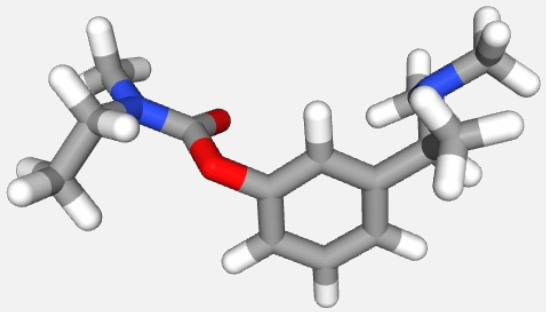
Github Link

APPENDIX

MOTIVATION UNIQUE CHALLENGES IN MOLECULE LANGUAGE MODEL

Challenge 2. Different Expertise with Diverse Interests

Molecule: Rivastigmine



FDA Pharm Classes

Main Interest
Approved indication of an active moiety that scientifically valid and clinically meaningful

Description for Rivastigmine
This molecule is a Cholinesterase Inhibitor. The mechanism of action of rivastigmine is as a Cholinesterase Inhibitor.

LiverTox

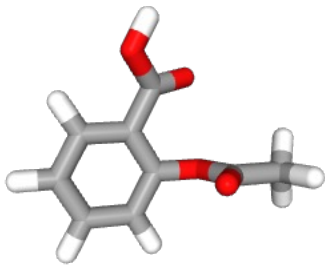
Main Interest
Injury attributable to prescription and nonprescription medications

Description for Rivastigmine
This molecule is an oral acetylcholinesterase inhibitor used for therapy of Alzheimer disease.

EXPERIMENTS ZERO-SHOT CROSS-MODAL RETRIEVAL

Task description

Choosing an appropriate description from several alternatives for a specific molecule (Given Molecule)
or retrieving the molecule that aligns with particular description (Given Text)

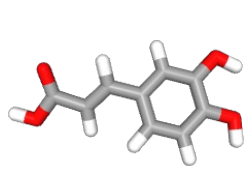


Aspirin

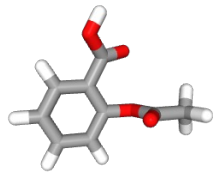
- ✗ This molecule is the sodium salt form of benzylpenicillin
- ✓ This molecule is a commonly used drug for the treatment of fever
- ✗ This molecule has been reported in Apis cerana.
- ✗ This molecule is a semi-synthetic penicillin antibiotic.

Given Molecule

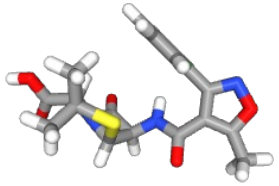
This molecule is a commonly used drug for the treatment of fever



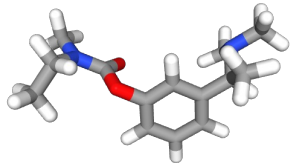
✗ Caffeic Acid



✓ Aspirin



✗ Cloxacillin



✗ Rivastigmine

Given Text

EXPERIMENTS ZERO-SHOT CROSS-MODAL RETRIEVAL

Task description

Choosing an appropriate description from several alternatives for a specific molecule (Given Molecule)
or retrieving the molecule that aligns with particular description (Given Text)

	SMILES	Graph	Given Molecule @ 20			Given Text @ 20		
			Descr.	Pharma.	ATC	Descr.	Pharma.	ATC
Single Encoder								
MolT5	✓	✗	5.06 (0.44)	6.80 (0.28)	6.48 (0.25)	6.66 (2.02)	6.02 (0.57)	6.10 (0.09)
BioT5	✓	✗	6.47 (0.13)	7.42 (0.52)	7.71 (0.16)	6.02 (0.42)	7.36 (0.13)	6.78 (0.45)
KV-PLM	✓	✗	42.28 (3.29)	36.84 (0.53)	30.21 (0.40)	45.64 (2.51)	37.93 (0.62)	33.22 (0.40)
Separate Encoder								
MoMu	✗	✓	**97.39** (0.19)	77.82 (0.54)	51.34 (0.37)	96.84 (0.17)	77.05 (0.28)	47.68 (0.34)
MoleculeSTM	✓	✗	96.70 (0.35)	77.28 (0.94)	52.36 (0.29)	96.22 (0.29)	75.01 (0.49)	50.01 (0.40)
MoleculeSTM	✗	✓	95.87 (1.87)	79.21 (0.75)	52.70 (0.75)	95.82 (0.37)	77.15 (0.74)	48.54 (0.49)
AMOLE	✗	✓	96.48 (2.94)	**81.46** (0.60)	**54.76** (0.57)	**97.20** (0.26)	**80.11** (0.42)	**51.47** (0.56)

Observations

- MoMu (Random augmentation of molecules) exhibit the lowest performance
→ Random augmentation of molecules can generate the molecules that are chemically invalid
→ Even worse than MoleculeSTM (No augmentation of molecules)

- AMOLE achieves the best results in five out of six datasets

- Performance gap widens with increasingly challenging tasks (Description < Pharm < ATC)