AIRI

Do LMs dream of molecule structures?

Lost in Translation: Chemical LMs and the Misunderstanding of Molecule Structures

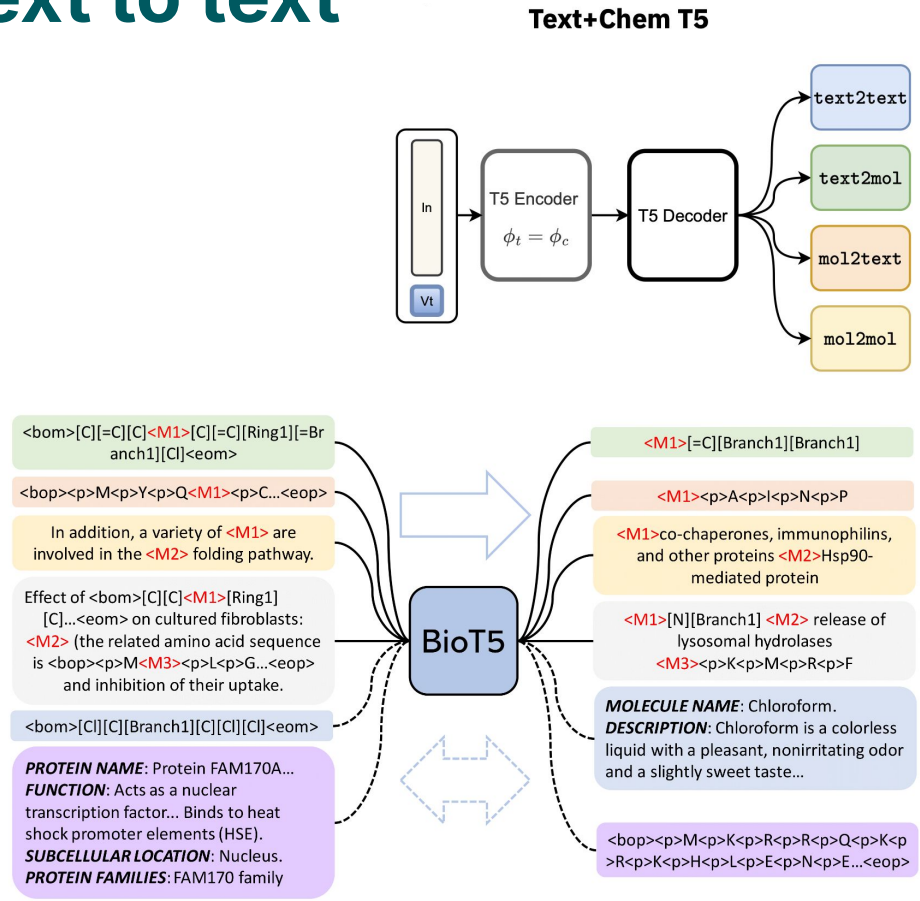
Veronika Ganeeva
Andrey Sakhovskiy
Kuzma Khrabrov
Artur Kadurin
Andrey Savchenko
Elena Tutubalina

Language Models: from text to text

LMs are Transformer-like architecture models and used to seq2seq tasks, when input and output are both texts.

Chemistry provide some text-based tasks like molecule captioning or molecular reaction result prediction.

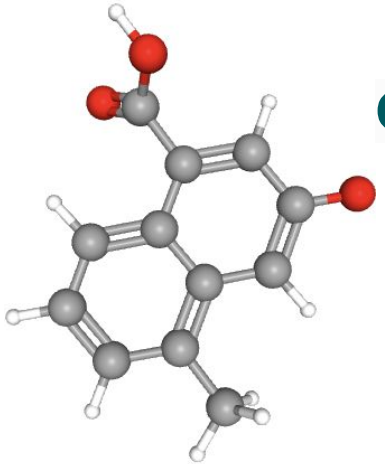
Textual representations of molecule structures allow to use LMs for chemical tasks: MolT5 (Edwards et al., 2022) Text+Chem T5 (Christofidellis et al., 2023)



SMILES: from molecule to text

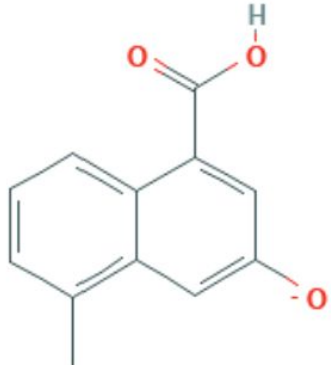
SMILES – best known string-based molecular representations.

Novel cross-domain LMs are pre-trained on both chemical and textual data for chemical tasks.



CC1=C2C=C(C=C(C2=CC=C1)C(=O)O)[O-]

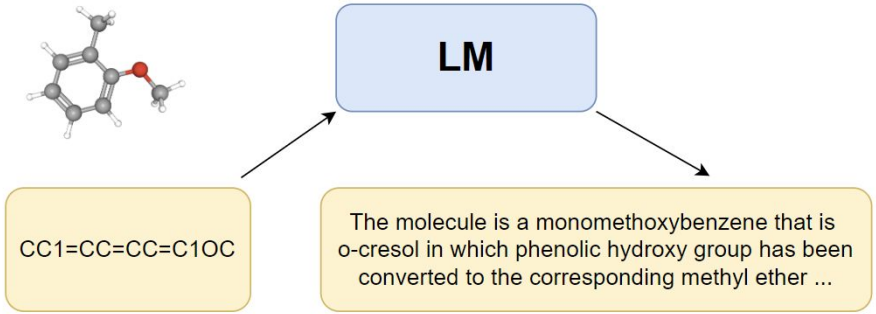
3-Hydroxy-5-methyl-1-naphthoate



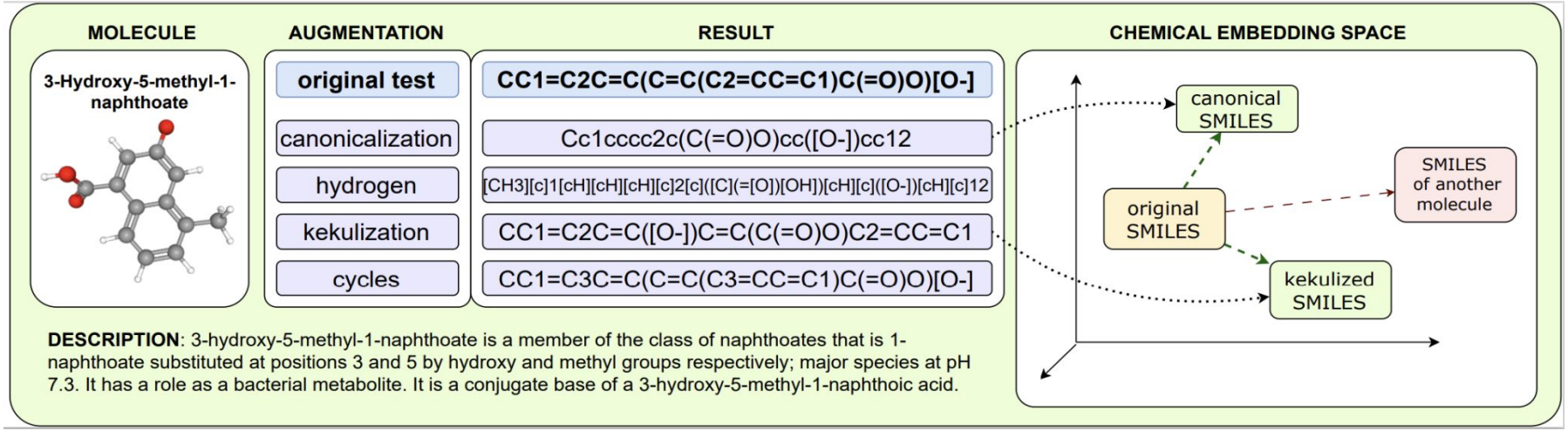
Do LMs reconstruct molecule structure from SMILES?

Do ChemLMs learn relationships within symbolic representations of molecular structures, enabling them to differentiate structures?

Molecule representation in LMs is crucial for enhancing chemical understanding. The valuation of chemical LMs is often conducted through downstream tasks that do not directly assess knowledge of chemistry

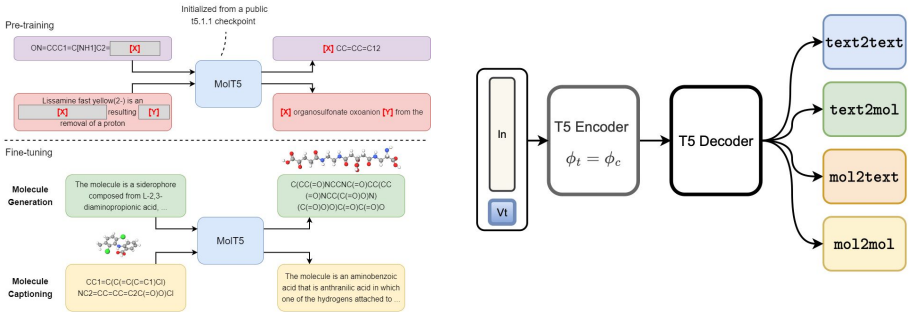


AMORE: evaluation framework



AMORE encodes original and augmented SMILES representations, calculates embedding distances, and assesses model performance based on top-1 accuracy, where the correct augmented SMILES is retrieved first

Models & Data



We augmented the ChEBI-20 dataset test part, which consists of 3,300 pairs of molecule description and filtered qm9 dataset, which consists of 990 isomeric molecules...

...and evaluate best known ChemLMs:
MolT5 (Edwards et al., 2022)
Text+Chem T5 (Christofidellis et al., 2023), etc

Model	Domain	# Params
Text+Chem T5-standard	Cross	220M
Text+Chem T5-augm	Cross	220M
MolT5-base	Cross	220M
MolT5-large	Cross	770M
SciFive	Text	220M
PubChemDeBERTa	Chem	86M
ChemBERT-ChEMBL	Chem	6M
ChemBERTa	Chem	125M
BARTSmiles	Chem	400M
ZINC-RoBERTa	Chem	102M
ZINC-GPT	Chem	87M

Carl Edwards, ChengXiang Zhai, and Heng Ji. Text2mol: Cross-modal molecule retrieval with natural language queries. EMNLP 2021
Carl Edwards et al. Translation between molecules and natural language. EMNLP 2022
Dimitrios Christofidellis et al. Unifying molecular and textual representations via multi-task language modelling. ICLR 2023

Results: molecule captioning

ChemLMs are not robust to augmentations and robustness to different augmentations varies. Captioning quality **is consistent with AMORE**
(CHEBI-20 dataset, Text+Chem T5 and MolT5)

Augmentation ⟶	canon			hydro		
Metrics	**Acc@1**	**ROUGE2**	**METEOR**	**Acc@1**	**ROUGE2**	**METEOR**
Text+Chem T5-standard	63.03	0.381	0.515	5.46	0.187	0.314
Text+Chem T5-augm	60.64	0.377	0.514	5.61	0.201	0.336
MolT5-base	42.88	0.315	0.450	2.36	0.199	0.329
MolT5-large	46.94	0.390	0.532	2.7	0.174	0.317
Augmentation ⟶	**kekul**			**cycles**		
Metrics	**Acc@1**	**ROUGE2**	**METEOR**	**Acc@1**	**ROUGE2**	**METEOR**
Text+Chem T5-standard	76.76	0.413	0.574	96.7	0.483	0.600
Text+Chem T5-augm	77.09	0.410	0.546	97.18	0.458	0.581
MolT5-base	62.76	0.333	0.475	90.94	0.417	0.540
MolT5-large	59.7	0.405	0.546	98.21	0.477	0.603

High metrics
Less changes
on SMILES

cycles →
kekule →
canonical →
hydrogen

Low metrics
More changes
on SMILES

Results: AMORE scores

AMORE allows to evaluate **different architectures**: encoders, decoders, encoder-decoders: Text+Chem T5, MolT5, ZINC-GPT, ChemBERTa, etc.

Model	Canon		Hydro		Kekul		Cycle	
	Acc@1	Acc@5	Acc@1	Acc@5	Acc@1	Acc@5	Acc@1	Acc@5
Cross-modal models								
Text+Chem T5-standard	63.03	82.76	5.46	10.85	76.76	92.03	96.7	99.82
Text+Chem T5-augm	60.64	82.79	5.61	12.64	77.09	92.06	97.18	99.7
MolT5-base	55.64	59.79	5.97	7.27	62.76	80.52	90.94	97.18
MolT5-large	46.94	63.58	2.36	5.06	59.7	75.84	98.21	100
Unimodal models								
BARTSmiles	25.76	38.09	1.21	2.15	39.03	54.97	61.67	71.24
ZINC-GPT	23.85	33.85	0.85	1.64	35.09	48.45	75.3	85.03
SciFive	29.73	44.94	2.58	4.64	48.21	68.15	98.48	100
PubChemDeBERTa	32.79	48.09	2.15	4.33	53.55	73.15	96.39	99.45
ChemBERT-ChEMBL	26.06	37.79	1.73	3.3	37.7	54.91	79.55	87.03
ChemBERTa	26.61	40.12	1.09	2.3	44.18	65.42	92.58	98.42
ZINC-RoBERTa	23.33	33.61	0.97	2.39	33.09	46.97	90.61	97.48

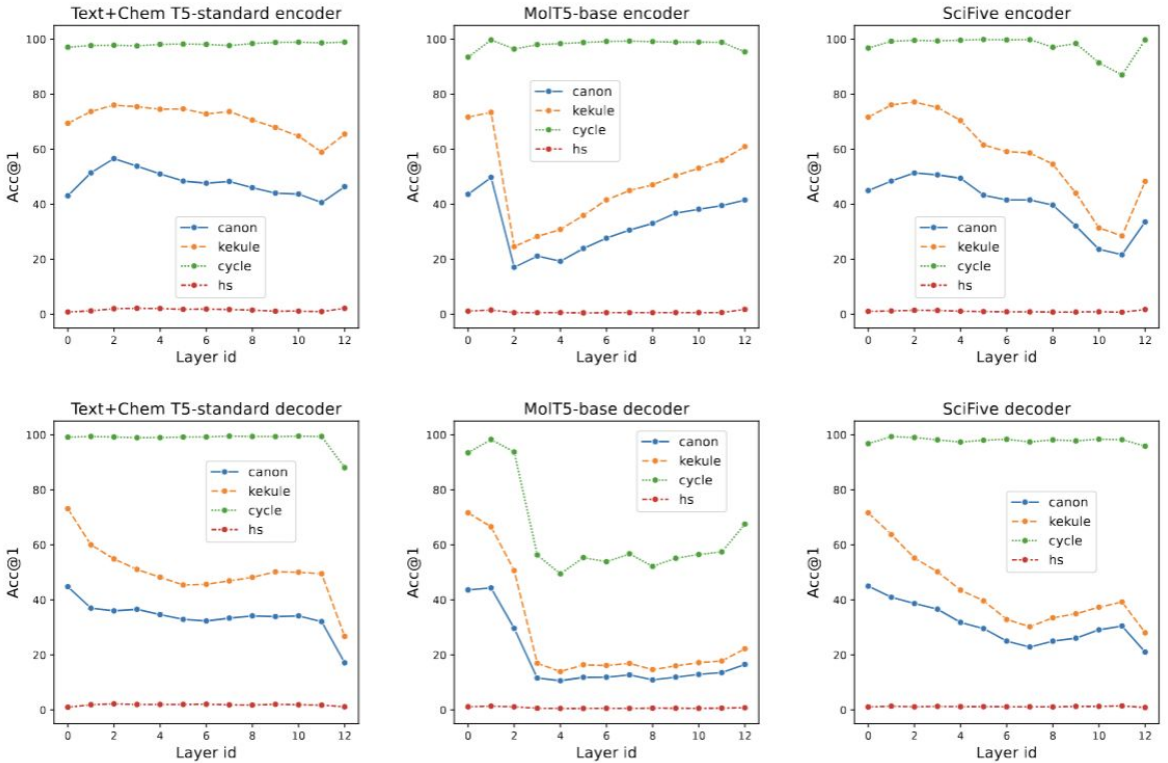
High metrics
Less changes
on SMILES

cycles →
kekule →
canonical →
hydrogen

Low metrics
More changes
on SMILES

Results: hidden states

Representation **robustness on model layers correlates** across different augmentations. Top-1 retrieval accuracy (Acc@1) on CheBI-20 dataset is calculated for hidden representations for different layers of LMs

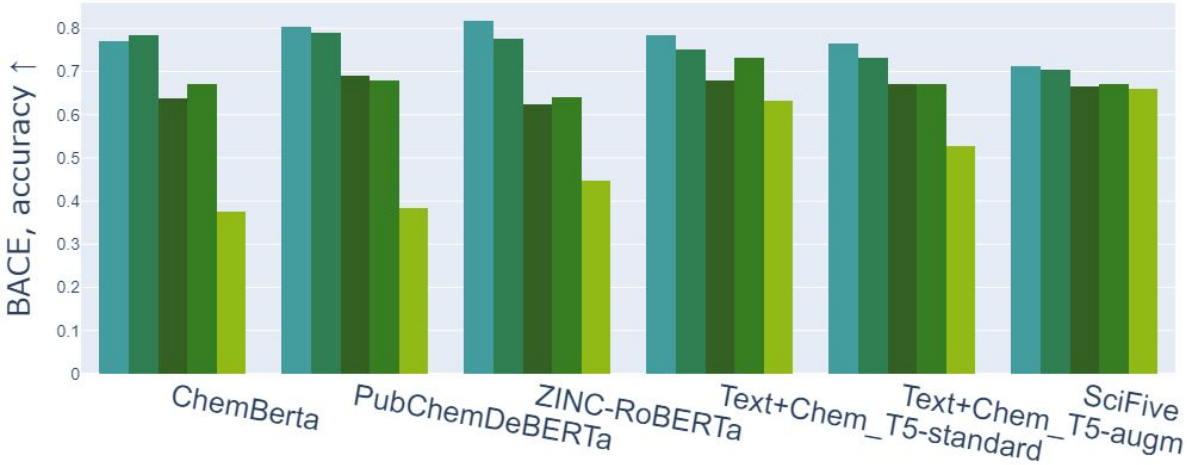


Results: MoleculeNet benchmark

BACE task (MoleculeNet): qualitative (binary label) binding results for a set of inhibitors of human β-secretase 1 (BACE-1).

Augmented SMILES **lead to degraded performance** on chemical tasks, robustness to different augmentations varies, range of robustness repeats.

Performance (BACE task) on five test sets: orig, cycle, canon, kekul, hydro



Summary

- We introduced **AMORE: flexible framework** for ChemLM evaluation
- We shows that **captioning quality is consistent** with AMORE
- Augmented SMILES **lead to degraded** performance on **chemical tasks**
- Representation **robustness on model layers correlates** across different augmentations

Datasets and code are publicly available