

Chemical Language Models Have Problems with Chemistry

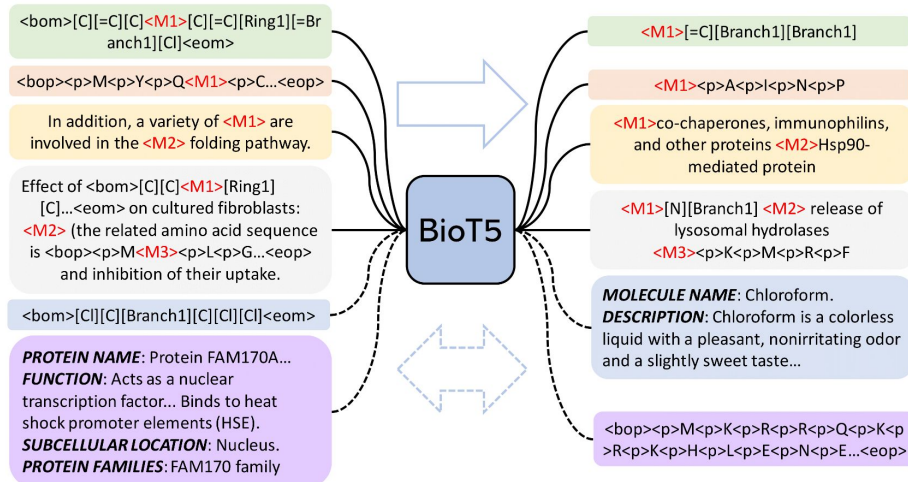
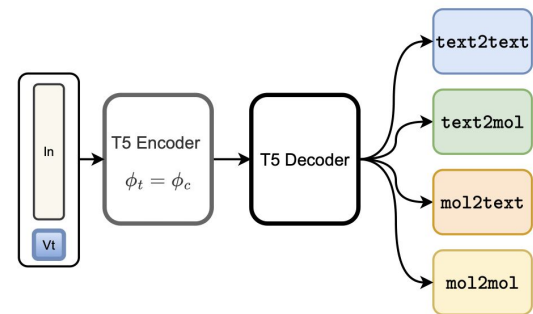
*Do LMs dream of
molecule structures?*

Veronika Ganeeva
Kuzma Khrabrov
Artur Kadurin
Andrey Savchenko
Elena Tutubalina

Language Models: from text to text

- LMs are used to seq2seq tasks like machine translation
- Chemistry provide tasks like molecule description or molecular reaction result prediction
- Textual representations of molecule structures allow to use LMs for chemical tasks

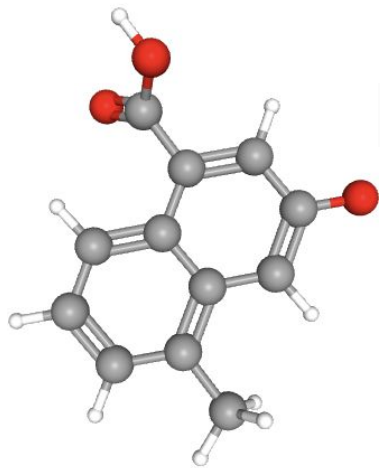
Text+Chem T5



SMILES: from molecule to text

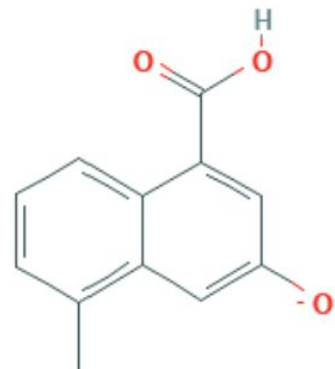
SMILES – best known string-based molecular representations.

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-]

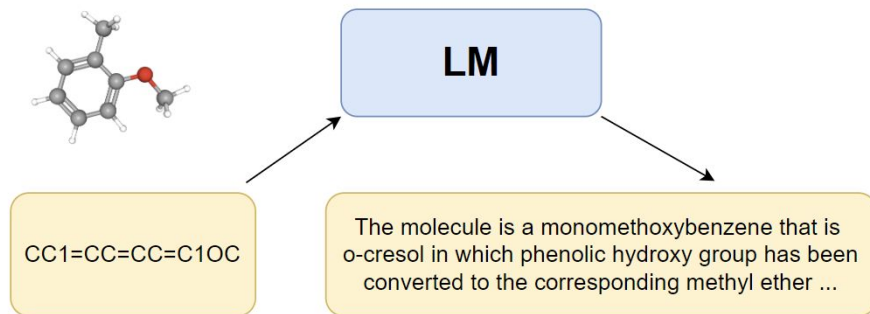
3-Hydroxy-5-methyl-1-naphthoate



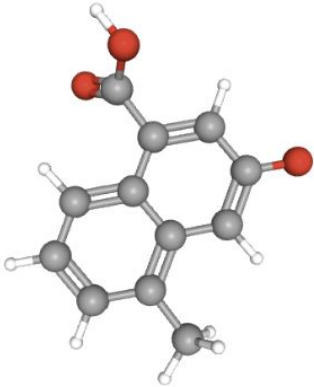
Do LMs reconstruct molecule structure from SMILES?

Or are they just guided by sequences of characters? Is there chemical knowledge?

To evaluate chemical knowledge of molecular structure in LMs for Chemistry we present probing tests. All tests are SMILES-based and transform SMILES into **equivalent variants** for the **same molecule structure**.



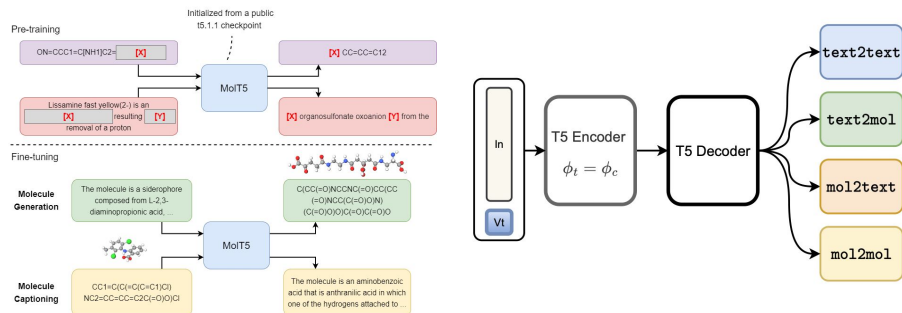
Probing tests

Molecule	Augmentation	Result
	original test	<chem>CC1=C2C=C(C=C(C2=CC=C1)C(=O)O)[O-]</chem>
	canonicalization	<chem>Cc1cccc2c(C(=O)O)cc([O-])cc12</chem>
	hydrogen	<chem>[CH3][c]1[cH][cH][cH][c]2[c]([C](=[O])[OH])[cH][c]([O-])[cH][c]12</chem>
	kekulization	<chem>CC1=C2C=C([O-])C=C(C(=O)O)C2=CC=C1</chem>
	cycles	<chem>CC1=C3C=C(C=C(C3=CC=C1)C(=O)O)[O-]</chem>
3-Hydroxy-5-methyl-1-naphthoate		

Models & Data

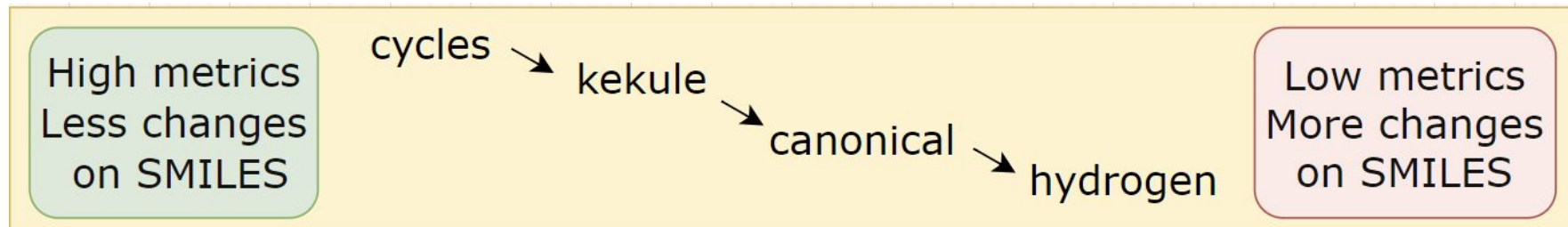
We augmented ChEBI-20 dataset test part, which consists of 3,300 pairs of molecule description...

...and evaluate best known cross-domain LMs: MolT5 (Edwards et al., 2022) and Text+Chem T5 (Christofidellis et al., 2023), that are pre-trained on both chemical and textual data and tasks



MODEL	Params	Fine-tune
MolT5-base	220M	ChEBI-20
MolT5-large	770M	ChEBI-20
Chem+TextT5 base	220M	ChEBI-20
Chem+TextT5 augm	220M	ChEBI-20, equal mixing strategy

Results



testset	MolT5-base		Chem+TextT5 _{base}		MolT5-large		Chem+TextT5 _{augm}	
	ROUGE-2	METEOR	ROUGE-2	METEOR	ROUGE-2	METEOR	ROUGE-2	METEOR
original	0.48	0.58	0.49	0.60	0.51	0.61	0.54	0.62
canonical	0.31	0.45	0.38	0.52	0.39	0.53	0.38	0.51
hydrogen	0.19	0.32	0.19	0.31	0.17	0.32	0.20	0.34
kekule	0.33	0.48	0.41	0.57	0.41	0.55	0.41	0.54
cycles	0.42	0.54	0.48	0.60	0.57	0.60	0.46	0.58

Summary

- We introduced **novel probing tasks** with chemistry LMs.
- **State-of-the-art chemical language models are vulnerable to changes** in molecule representations.
- All **changes in symbolic representation** have proven to **cause a decline** in performance.
- Extent of this decline seems to be dictated by **language processing rather than** the underlying **understanding of chemistry**.
- This new information will allow the scientific community to better **understand the domain-specific capabilities** achieved by novel cross-domain LMs.

Datasets and code are publicly available

