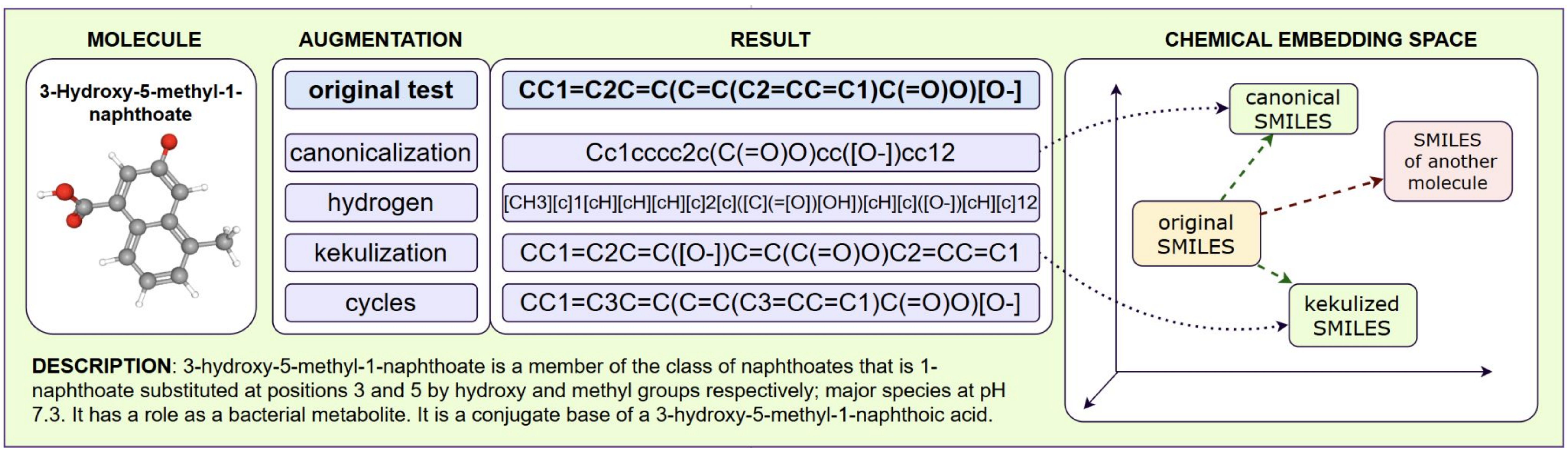
Lost in Translation: Chemical LMs and the Misunderstanding of Molecule Structures

V. Ganeeva, A. Sakhovskiy, K. Khrabrov, A. Savchenko, A. Kadurin, E. Tutubalina



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

Cc1cccc2c(C(=O)O)cc([O-])cc12

[CH3][c]1[cH][cH][cH][c]2[c]([C](=[O])[OH])[cH][c]([O-])[cH][c]12

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

CC1=C3C=C(C=C(C3=CC=C1)C(=O)O)[O-]

DESCRIPTION: 3-hydroxy-5-methyl-1-naphthoate is a member of the class of naphthoates that is 1-naphthoate substituted at positions 3 and 5 by hydroxy and methyl groups respectively; major species at pH 7.3. It has a role as a bacterial metabolite. It is a conjugate base of a 3-hydroxy-5-methyl-1-naphthoic acid.

Fig.1 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

Problem: Molecule representation in LMs is crucial for enhancing chemical understanding. The valuation of ChemLMs is conducted through downstream tasks that don't directly assess knowledge of chemistry

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

Table 1. AMORE: Acc@1, Acc@5 on the ChEBI-20 data

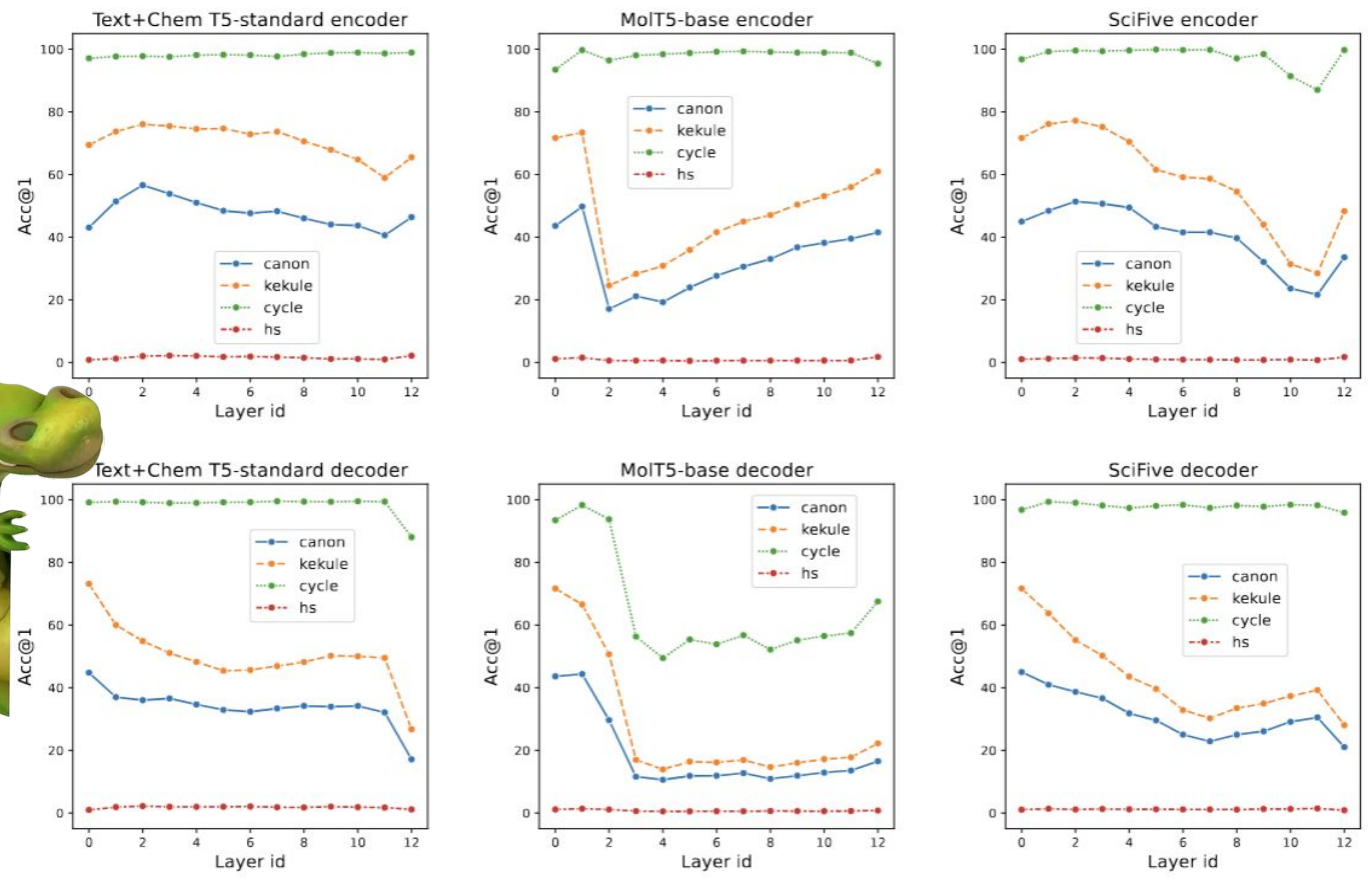
Model	Canon		Hydro		Kekul		Cycle	
	Acc@1	Acc@5	Acc@1	Acc@5	Acc@1	Acc@5	Acc@1	Acc@5
Cross-modal models								
Text+Chem T5-standard	63.03	82.76	5.46	10.85	76.76	92.03	96.7	99.82
Text+Chem T5-augm	60.64	82.79	5.61	12.64	77.09	92.06	97.18	99.7
MolT5-base	55.64	59.79	5.97	7.27	62.76	80.52	90.94	97.18
MolT5-large	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

Fig.2 Top-1 retrieval accuracy (Acc@1) on CheBI-20 dataset calculated for hidden representations for different layers of LMs



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

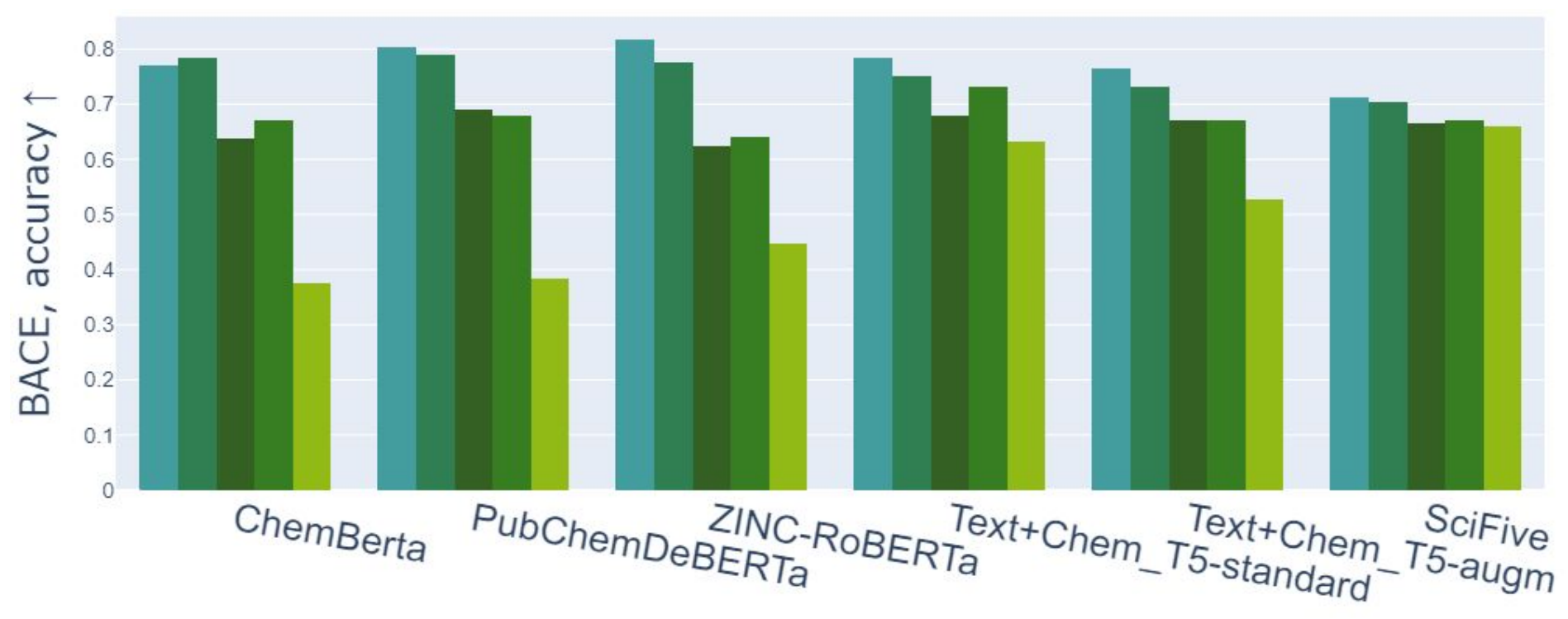
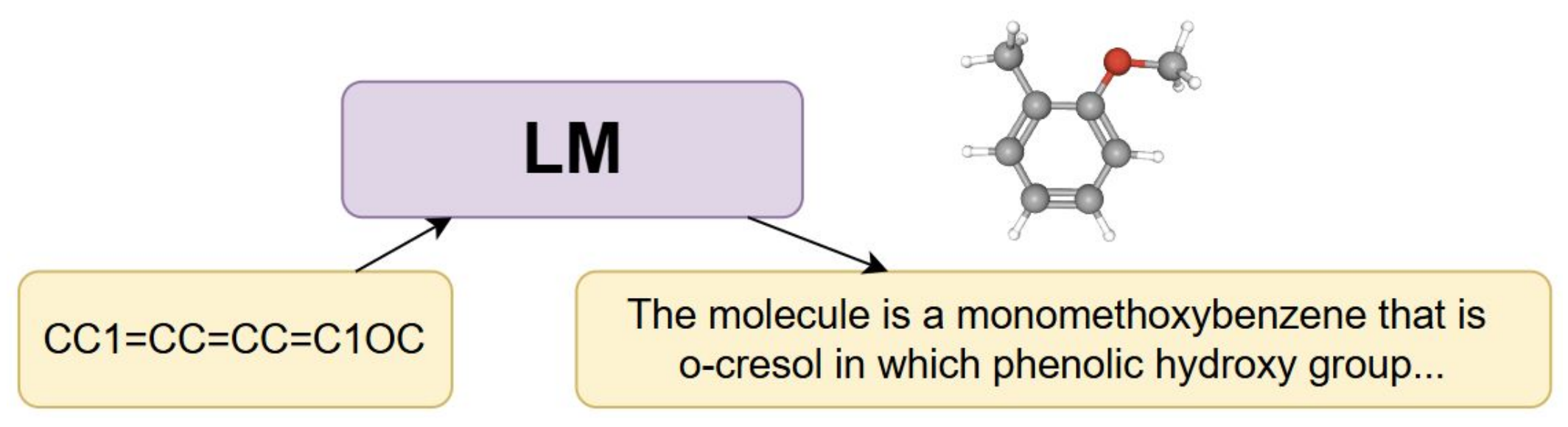


Fig.3 Performance (BACE task) on five test sets: **orig, cycle, canon, kekul, hydro**

Framework AMORE: embeddings distance + augmentations of the same molecule (Fig. 1)

AMORE allows to evaluate different architectures (results in Tab. 1) and compare robustness on model hidden layers (Fig. 2)

Table 2. Acc@1 & METEOR on the molecule captioning task (CHEBI-20 test)


LM
CC1=CC=CC=C1OC
The molecule is a monomethoxybenzene that is o-cresol in which phenolic hydroxy group...

Augmentation ⟶	canon		hydro		kekul		cycles	
Metrics	Acc@1	METEOR	Acc@1	METEOR	Acc@1	METEOR	Acc@1	METEOR
Text+Chem T5-standard	63.03	0.515	5.46	0.314	76.76	0.574	96.7	0.600
Text+Chem T5-augm	60.64	0.514	5.61	0.336	77.09	0.546	97.18	0.581
MolT5-base	42.88	0.450	2.36	0.329	62.76	0.475	90.94	0.540
MolT5-large	46.94	0.532	2.7	0.317	59.7	0.546	98.21	0.603

Results:

- ChemLMs are **not robust** to augmentations
- **Robustness** to augmentations **varies**
- Augmented SMILES **lead to degraded performance** on chemical tasks
- Captioning quality is **consistent with AMORE**
- Representation **robustness on model layers correlates** across augmentations
- ChemLMs **benefit** from **cross-modality**

Datasets and code are publicly available:

Questions: ganeeva@airi.net