Lost in Translation: Chemical Language Models and the Misunderstanding of Molecule Structures

Veronika Ganeeva[1]*, Andrey Sakhovskiy[2,3]*, Kuzma Khrabrov[1], Andrey Savchenko[4,5,6], Artur Kadurin[1,6], Elena Tutubalina[1,2,4,6]

[1]AIRI [2]Sber AI [3]Skoltech [4]HSE University
[5]Sber AI Lab [6]ISP RAS Research Center for Trusted Artificial Intelligence

Correspondence: {ganeeva, tutubalina}@airi.net

Abstract

The recent integration of chemistry with natural language processing (NLP) has advanced drug discovery. Molecule representation in language models (LMs) is crucial in enhancing chemical understanding. We propose **A**ugmented **Mo**lecular **Re**trieval (♡AMORE), a flexible zero-shot framework that assesses trustworthiness of Chemical LMs of different natures: trained solely on molecules for chemical tasks and on a combined corpus of natural language texts and string-based structures. The framework relies on molecule augmentations that preserve an underlying chemical, such as kekulization and cycle replacements. We evaluate encoder-only and generative LMs by calculating a metric based on the similarity score between distributed representations of molecules and their augmentations. Our experiments on ChEBI-20 and QM9 benchmarks show that these models exhibit significantly lower scores than graph-based molecular models trained without language modeling objectives. Augmentation of SMILES representations leads to decreased performance on chemical property prediction tasks in the MoleculeNet benchmark. Additionally, our results on the molecule captioning task for cross-domain models, MolT5 and Text+Chem T5, demonstrate that our representation-based evaluation metrics significantly correlate with the classical text generation metrics like ROUGE and METEOR.

1 Introduction

Drawing inspiration from the progress of Transformer-like architectures in NLP (Vaswani et al., 2017), the drug discovery community has embraced state-of-the-art molecular-generation methodologies. This includes leveraging LM-based approaches such as ChemBERTa, T5Chem, ChemFormer, and BARTSmiles (Chithrananda et al., 2020; Lu and Zhang, 2022; Irwin et al., 2022; Chilingaryan et al., 2022). SMILES (Simplified Molecular Input Line Entry System) (Weininger, 1988), a commonly employed molecular representation, enables the use of molecules in a string-based format. These single-domain models are usually pre-trained on large SMILES datasets like ZINC-15 (Sterling and Irwin, 2015), then fine-tuned for downstream tasks like reaction and property prediction on datasets like USPTO (Lowe, 2012, 2017) and MoleculeNet (Wu et al., 2018a).

Recently, LMs like MolT5 (Edwards et al., 2022a) and Text+Chem T5 (Christofidellis et al., 2023) have been introduced to integrate chemical and linguistic knowledge. These models were pre-trained on chemical and textual data, e.g., the large C4 (Raffel et al., 2020) and ZINC-15, and fine-tuned on cross-domain tasks like molecule captioning. However, the evaluation of downstream tasks does not directly assess knowledge of chemistry, such as understanding various representations of the same molecular chemical structure. In this work, we seek to answer the following research question (RQ):

> Do chemical language models (ChemLMs) learn patterns and relationships within symbolic representations of molecular structures during pre-training, enabling them to differentiate molecular structures based on learned patterns?

Our main contribution is the novel unified approach (Fig. 1) to verify if ChemLMs have effectively grasped the fundamental rules on the construction of molecular representations, such as SMILES. Our hypothesis posits that augmentation should not significantly alter the similarity score between distributed representations of molecules and their augmented versions. To address the RQ, we conduct experiments using BERT-based (Devlin et al., 2019), GPT-based (Brown et al., 2020), and T5-based (Raffel et al., 2020) models.

*These authors contributed equally to this work. The order of author names was randomly determined.

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 12994–13013
November 12-16, 2024 ©2024 Association for Computational Linguistics

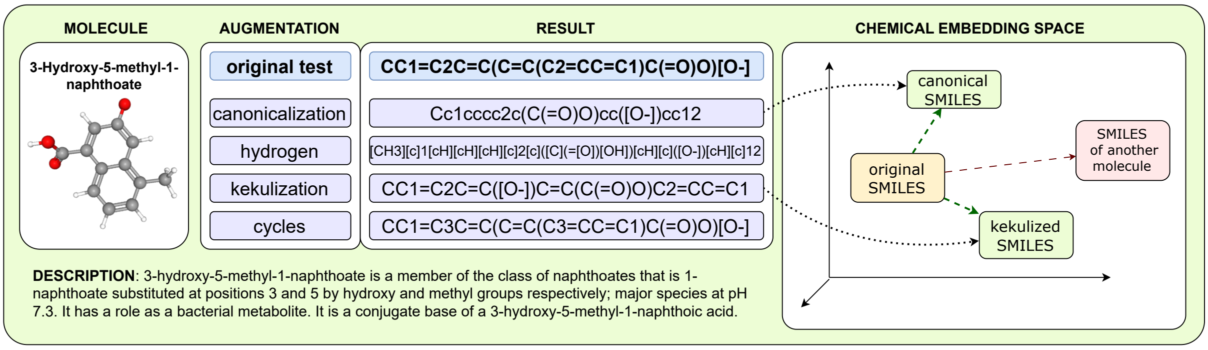
MOLECULE
AUGMENTATION
RESULT
CHEMICAL EMBEDDING SPACE
3-Hydroxy-5-methyl-1-naphthoate
original test
CC1=C2C=C(C=C(C2=CC=C1)C(=O)O)[O-]
canonicalization
Cc1cccc2c(C(=O)O)cc([O-])cc12
hydrogen
[CH3][c]1[cH][cH][cH][c]2[c]([C](=[O])[OH])[cH][c]([O-])[cH][c]12
kekulization
CC1=C2C=C([O-])C=C(C(=O)O)C2=CC=C1
cycles
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.
canonical SMILES
original SMILES
SMILES of another molecule
kekulized SMILES

Figure 1: Our evaluation involves generating augmented molecules for all molecules in a dataset using one of four augmentation procedures. By encoding original molecules and augmented SMILES representations and calculating distances between embeddings, the study determines model performance based on top-1 accuracy, where the correct augmented SMILES must be retrieved at the top rank.

Our work is designed to provide insights into the extent to which ChemLMs can discern molecular structures and the impact of augmentation on their performance. To summarize, our main contributions are as follows:

- We propose **A**ugmented **Mo**lecular **Re**trieval (♡AMORE)[1], a novel framework for quality assessment of chemical language models. It relies on augmentations of molecular SMILES string representations that are known to produce alternative representations without changing an underlying molecule. Unlike supervised fine-tuning-based methods for chemical LM quality assessment adopted from NLP, our framework adopts known chemical facts to perform a fully unsupervised evaluation in a zero-shot setting.

- Extensive evaluation has revealed that embedding spaces of the existing state-of-the-art unimodal and cross-modal chemical LMs are not robust to four SMILES augmentation types known to be identity transformations in terms of the underlying molecules.

- We show that the state-of-the-art ChemLMs fail on a wide range of molecule understanding tasks when an augmented representation of the same molecule is passed. The explored tasks include molecule captioning, molecular binary and multi-class classification, and molecular property prediction.

2 ♡AMORE: Augmented Molecular Retrieval

In this section, we introduce AMORE, a flexible embedding-based evaluation framework for Language Model analysis in the chemistry domain.

Methodology. Our evaluation metrics are built on distributed representations of molecules and their augmentations. Let X_1 denote the dataset comprising original representations of molecules, represented as $x_1, x_2, \ldots, x_n$. Through SMILES augmentation, we generate the X_1' dataset, containing augmented representations of the same molecules, represented as $x_1', x_2', \ldots, x_n'$. In each experiment, a model encodes the augmented SMILES representations of molecules. Let $e(x_i)$ represent the embedding of SMILES x_i from the original dataset, and $e(x_j')$ represent the embedding of the augmented SMILES x_j' from the augmented dataset, where i, j denote indices corresponding to molecules. The distance between embeddings $e(x_i)$ and $e(x_j')$ is computed using a distance metric such as Euclidean distance. Suppose we retrieve the closest SMILES representation to the original one x_i among the augmented ones x_j'. When the nearest molecule from the augmented dataset is not an augmentation of the original SMILES $j \neq i$, this indicates that a ChemLM does not recognize the same chemical structure within the augmented textual embedding space. The task formulation strongly resembles entity linking (EL) where the goal is to map an extracted entity to the most relevant concept from a pre-defined vocabulary or knowledge graph. Thus, we adopt the standard EL evaluation methodology (Tutubalina et al., 2020; Sakhovskiy et al., 2023, 2024) and use the top-k accuracy as the evaluation metric: $\mathrm{Acc@k} = 1$ if

[1] AMORE is available at https://github.com/ChemistryLLMs/AMORE

the correct augmented SMILES is retrieved at the rank $\leq k$, otherwise $\mathrm{Acc@k} = 0$; in our case, k=1. In addition, we compute the Mean Reciprocal Rank (MRR) metric. This ranking metric can get a better sense of the performance degradation with the augmented SMILES strings since it reflects the average ranking of true molecule (Radev et al., 2002).

The practical objective of our approach is to compare embeddings for different textual representations of the same molecule structure. We use the fast nearest neighbor search library FAISS (Johnson et al., 2019) that is efficient in a large-scale setting. Our methodology's theoretical implications lie in understanding how efficiently ChemLMs reconstruct molecule structures from the textual representations provided to them.

Augmentation Procedures. We follow four popular augmentations from (Ganeeva et al., 2024), where the authors showed that augmentations led to a decrease in ROUGE scores (Lin, 2004) in the molecule captioning task when evaluated two cross-domain T5 models, Text+Chem T5 and MolT5. It is important to note that the use of machine translation metrics restricts the number of models that can be compared, as these cross-domain architectures require natural language tokens. We adopt the following SMILES-based augmentation procedures: 1. **Canonicalization**: we transform SMILES strings into a standardized RDKit string (Bento et al., 2020; Greg et al., 2022), reducing ambiguity and facilitating accurate molecule comparisons. 2. **Hydrogen**: Hydrogen is commonly omitted from SMILES, as restoring expected atom connections is straightforward. However, molecule properties depend on 3D chirality, which can depend on hydrogen position. Explicitly including hydrogen in SMILES allows a more comprehensive representation, especially for hydrogen bonding, stereochemistry, and reaction mechanisms. 3. **Kekulization**: Aromaticity is an essential concept in organic chemistry, influencing molecular stability, reactivity, and spectroscopic properties. This involves transforming a SMILES string into a Kekulized SMILES string, where the aromatic π-electrons are static between every second carbon; 4. **Cycles**: In chemical graph theory, cycles (or rings) play a fundamental role in characterizing molecular structure and properties. Valid replacement of cycle numerical identifiers with other random numbers allows for testing the robustness of models in recognizing cyclic structures and their connectivity, providing insights into their ability to handle diverse molecular topologies.

The key property of the four augmentations listed above is that a resulting augmented SMILES represents the same molecule as the original non-augmented one. Intuitively, these augmentations can be seen as identity transformations on molecules (i.e., x_i and x'_i are two different strings representing the same underlying chemical). For instance, the canonical SMILES for methane is "C", while the full version is "[CH4]" (carbon atom is connected to four hydrogen atoms). As in organic chemistry, a carbon atom C is implied to be connected to hydrogen atoms by default; hydrogen atoms H are usually omitted for brevity. For more examples of augmented SMILES, see Appx. A.

Overall, our ♡AMORE framework can be briefly summarized as follows:

1. Take a set $X = (x_1, x_2, \ldots, x_n)$ consisting of n molecular representations;

2. Apply a transformation f to obtain a set of augmented molecular representations $X' = (x'_1, x'_2, \ldots, x'_n)$, where $x'_i = f(x_i)$. The only constraint introduced for f is that it should not change an underlying chemical. We execute all augmentations through RDKit, a widely recognized methodology within the chemistry domain (Bento et al., 2020). As in this work, we focus on textual molecular representations, we can think of x_i and x'_i as being **synonyms**.

3. For each $x_i \in X$ and $x'_j \in X'$ obtain their vectorized representations $e(x_i)$ and $e(x'_j)$, respectively.

4. Evaluate the vectorized representations in a retrieval task: given an embedding $e(x_i)$, a model should be able to retrieve an embedding $e(x'_i)$ of augmented x'_i.

The augmentation vectors are in a similar embedding space, allowing distance measurement between original and augmented molecules. The better a model performs the AMORE retrieval task, the more robust it is to the transformation f, indicating the model **knows** f is a **mapping between synonymous** representations.

Datasets. Our evaluation strategy relies on two popular datasets: (i) a ChEBI-20 test set (Edwards et al., 2021) and (ii) a subset from the QM9 (Ruddigkeit et al., 2012; Ramakrishnan et al., 2014) (further called **Isomers**), consisting of isomers of C9H12N2O. We select these datasets for the following reasons: 1. Utilizing the ChEBI-20 test set, which comprises approximately 3k molecule-description pairs, allows for comparisons with metrics such as ROUGE (Lin, 2004) and METEOR (Banerjee and Lavie, 2005) in the molecule captioning task. The ChEBI-20 train set was used to train cross-domain ChemLMs. Hence, we follow the recent papers (Christofidellis et al., 2023; Edwards et al., 2022a), which use ChEBI-20 for benchmarking on molecule captioning tasks. 2. The ChEBI-20 dataset comprises molecular structures that translate into SMILES strings of varying lengths. This diversity in sequence length and symbol sets could potentially impact the mean characteristics of accurately identified results. 3. Furthermore, some molecules in the ChEBI-20 dataset may not be suitable for augmentation using our proposed methods. For instance, cycle renumbering relies on aromatic hydrocarbons, which are absent in non-organic compounds. This limitation may affect the comprehensiveness of our evaluation.

Due to these reasons, it is essential to complement the evaluation with datasets that mitigate these weaknesses. Therefore, we have selected molecules from the QM9 dataset presented in the PubChem database (Kim et al., 2016). There are 3300 and 918 molecules in the ChEBI-20 test set and the Isomers set, respectively.

3 Models

For our experiments, we adopted various Transformer-based (Vaswani et al., 2017) molecular representation models, including encoder-only, encoder-decoder, and decoder-only architectures. All models are publicly available at HuggingFace. For more details, please see Appendix E.

Encoder-only A common approach is to train BERT-based encoders on unlabeled SMILES using objectives like Masked Language Modeling. We evaluate: (i) *PubChemDeBERTa* (Schuh et al., 2024), (ii) *ChemBERT-ChEMBL* (Zhang et al., 2022), (iii) *ChemBERTa* (Chithrananda et al., 2020), and (iv) *ZINC-RoBERTa* that are pre-trained on SMILES from various chemical databases, namely, PubChem (Kim et al., 2023) and ZINC (Sterling and Irwin, 2015). Some models, e.g., ChemBERT-ChEMBL and ChemBERTa, are known to be trained with augmented data.

Encoder-decoder We focus on two recent cross-modal T5-based (Raffel et al., 2020) for text-related chemical tasks: (i) *Text+Chem T5* (Christofidellis et al., 2023) and (ii) *MolT5* (Edwards et al., 2022b). We utilize base and large versions of *MolT5* and two base-sized versions of *Text+Chem T5*. Additionally, we employed a biomedical LM *SciFive* (Phan et al., 2021), a uni-modal textual T5-based model pre-trained on the general-domain C4 corpus and PubMed database.

Decoder-only As a decoder-only model, we adopt *ZINC-GPT* (Karl, 2024), a GPT-like (Radford et al., 2019) autoregressive language model trained on 480K SMILES from the ZINC database.

Graph Neural Networks The different family of models, so-called Graph Neural Networks (GNNs), treat SMILES inputs as graph objects and run the message-passing algorithm to produce nodes, vertices, and complete graph embeddings (Gilmer et al., 2020). The GNN performance can be considered as the upper limit for language models since GNNs are robust to the proposed augmentations (only edges' types in graph structures may change due to the chirality encoding). We used a GNN checkpoint available from the DGL-LifeSci framework (Li et al., 2021) to check this behavior.

4 Experimental Results

4.1 Molecule-augmentation retrieval

Our goal is to develop a unified framework to assess trustworthiness of chemical models, which is crucial for their application in chemical NLP tasks. In particular, we use the AMORE framework to empirically explore how well existing chemical language models deal with "synonymous" SMILES representations of the same molecule. Given an original SMILES x_i, we rank all augmented representations x'_j in terms of similarity between pooled representations $e(x_i)$ and $e(x'_j)$ obtained from a chemical LM. We assume that if a model retrieves an augmented x'_i of a higher rank given x_i, it is robust to the selected augmentation and is aware that the given augmentation is an identity transformation in molecules. E.g., the degraded performance suggests models struggle to distinguish non-augmented molecules from those with added hydrogen, implying they are not aware hydrogen is often omitted but implicitly present in SMILES. We use mean-pooled embeddings from Transformer layers as representations of SMILES.

Model	Canon			Hydro			Kekul			Cycle		
	Acc@1	Acc@5	MRR	Acc@1	Acc@5	MRR	Acc@1	Acc@5	MRR	Acc@1	Acc@5	MRR
Cross-modal models												
Text+Chem T5-standard	63.03	82.76	72.4	5.46	10.85	8.6	76.76	92.03	83.8	96.7	99.82	98.2
Text+Chem T5-augm	60.64	82.79	70.9	5.61	12.64	7.1	77.09	92.06	84.4	97.18	99.7	98.3
MolT5-base	55.64	59.79	50.9	5.97	7.27	5.5	62.76	80.52	70.9	90.94	97.18	93.8
MolT5-large	46.94	63.58	54.7	2.36	5.06	4.1	59.7	75.84	67.2	98.21	100	99.1
Unimodal models												
BARTSmiles	25.76	38.09	31.8	1.21	2.15	2.2	39.03	54.97	46.9	61.67	71.24	66.2
ZINC-GPT	23.85	33.85	28.8	0.85	1.64	1.5	35.09	48.45	41.7	75.3	85.03	80.1
SciFive	29.73	44.94	39.9	2.58	4.64	2.9	48.21	68.15	62.4	98.48	100	99.2
PubChemDeBERTa	32.79	48.09	40.3	2.15	4.33	3.6	53.55	73.15	62.9	96.39	99.45	97.9
ChemBERT-ChEMBL	26.06	37.79	32.2	1.73	3.3	2.8	37.7	54.91	46.1	79.55	87.03	83.2
ChemBERTa	26.61	40.12	33.3	1.09	2.3	2.1	44.18	65.42	54.1	92.58	98.42	95.3
ZINC-RoBERTa	23.33	33.61	33.2	0.97	2.39	1.7	33.09	46.97	45.5	90.61	97.48	69.2
Graph Neural Network model												
GNN	92.42	98.24	95.1	92.42	98.24	95.1	92.51	98.24	95.1	99.15	100.	99.6

Table 1: Top-1 / Top-5 accuracy (%) and Mean Reciprocal Rank (MRR) of ChemLMs and GNN for matching of distributed representations of molecules with their augmentations on the ChEBI-20 dataset.

Model	Canon			Hydro			Kekul			Cycle		
	Acc@1	Acc@5	MRR	Acc@1	Acc@5	MRR	Acc@1	Acc@5	MRR	Acc@1	Acc@5	MRR
Cross-modal models												
Text+Chem T5-standard	36.93	59.8	72.41	0.65	2.94	8.57	42.92	66.34	83.78	80.94	98.58	98.18
Text+Chem T5-augm	39	63.62	70.89	0.65	5.12	7.11	45.21	70.7	84.39	80.94	98.58	98.35
MolT5-base	29.96	44.55	37.62	0.54	3.16	2.65	36.17	51.96	44.32	76.36	92.37	83.52
MolT5-large	29.41	42.81	37.45	1.53	6.75	3.16	35.29	49.13	43.41	81.7	98.15	90.72
Unimodal models												
BARTSmiles	27.89	42.05	31.76	0	0.87	1.11	31.81	48.58	37.38	41.83	44.77	42.43
ZINC-GPT	24.18	36.17	32.03	0.44	1.31	0.97	27.45	43.03	37.69	55.12	68.52	64.41
SciFive	22	33.44	39.95	0	1.2	2.97	24.62	37.8	62.41	93.14	98.04	99.22
PubChemDeBERTa	26.69	38.13	31.96	0.22	0.65	0.99	31.59	44.88	37.8	87.36	94.99	90.82
ChemBERT-ChEMBL	23.64	34.86	31.52	0.98	3.38	1.34	27.12	39.54	36.23	37.15	39.76	65.99
ChemBERTa	25.93	36.6	31.69	0.65	2.94	1.74	29.3	41.72	36.46	50.98	60.13	80.49
ZINC-RoBERTa	28.76	42.27	36.48	0.65	1.85	1.33	33.12	49.35	42.61	50.76	56.86	84.64
Graph Neural Network model												
GNN	80.61	96.40	87.3	80.93	96.40	87.4	80.93	96.40	87.4	100.	100.	100.

Table 2: Top-1 / Top-5 accuracy (%) and Mean Reciprocal Rank (MRR) of ChemLMs and GNN for matching of distributed representations of molecules with their augmentations on the Isomers dataset.

Our results for matching distributed representations of molecules with their augmentations on CheBI-20 and Isomers datasets are presented in Table 1 and Table 2. Higher top-1/ top-5 accuracy and MRR indicate a model can recognize that varying SMILES representations correspond to the same molecule, i.e., robust to that type of augmentation.

Chemical LMs are not robust to SMILES augmentations The existing ChemLMs struggle to retrieve augmented SMILES for non-augmented ones indicating that they are unable to recognize synonymous SMILES variations. The finding lets us assume that pre-training on SMILES leads to memorization rather than an understanding of chemistry and a poor generalization. No model performs best on all augmentations and datasets, but retrieval is higher on the less complex ChEBI-20 dataset, possibly due to the transformation of short and non-aromatic molecules by our augmentations being less frequent.

Robustness to different types of augmentations varies significantly For all ChemLMs, augmentation ordering concerning retrieval accuracy remains consistent: the most challenging augmentation is explicit hydrogen addition, then transforming into RDKit canonical, kekulization, and cycle renumbering. Encoder-only PubChemDeBERTa, ChemBERTa, and ZINC-RoBERTa models are not far behind T5 models for cycle renumbering augmentation on ChEBI-20. Surprisingly, retrieval accuracy for hydrogen addition augmentation is extremely low for all models. On Isomers, all models have failed to surpass 1% accuracy. We believe that poor performance on hydrogen addition is caused by its absence in pre-training data of these models: hydrogen is always omitted whenever possible.

We note that from the construction, the output of chemistry GNNs depends only on the molecular

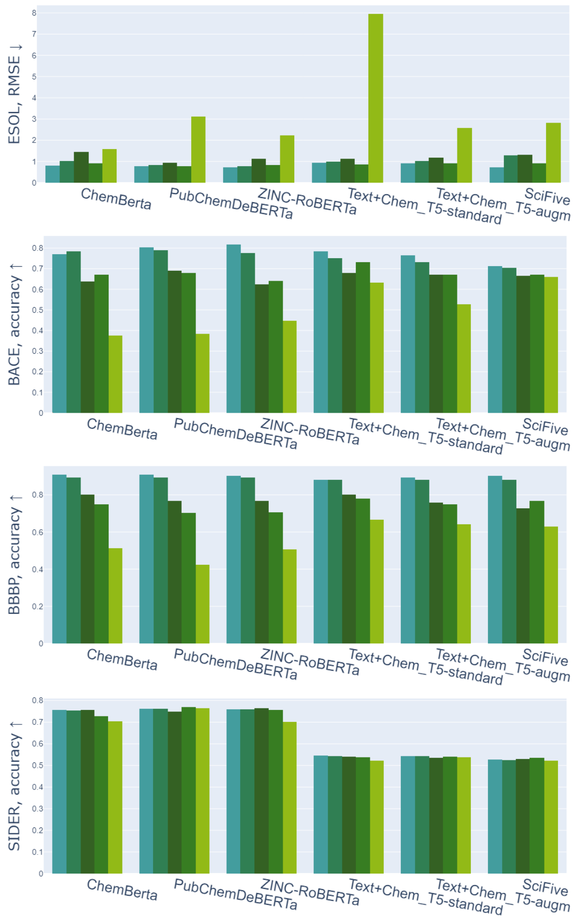

Figure 2: Performance on original and augmented MoleculeNet test sets, showing the impact of different data augmentation techniques on model performance across regression (ESOL), binary classification (BBBP, BACE), and multilabel classification tasks (SIDER).

graph structure, which can only be slightly changed by the proposed augmentations (only edge types change due to the chirality encoding). Thus, such models outperform ChemLMs on the AMORE metrics, as shown in Tab. 2. Though it remains questionable whether comparing GNNs and LMs is fair, we still find this comparison useful because, in general, we would like to probe chemical ML models of different natures. Indeed, GNNs do not solve the NLP tasks out of the box, so we should consider their metrics as the baseline for ChemLMs, where the latter may be augmented with the graph representations. A potential conclusion is incorporating parsing or graph representations into future ChemLMs for improvements.

Chemical LMs benefit from cross-modality For three augmentations except for cycle renumbering, cross-modal models (MolT5 and Text+Chem T5 variations) pre-trained on textual and chemical tasks yield higher retrieval accuracy consistently. Scores of Text+Chem T5-standard and Text+Chem T5-augm are, in most cases, higher than scores of other models. Interestingly, SciFive is the most robust to cycle renumbering on both datasets, even though it is pre-trained on texts only with no SMILES. The obtained top-1 accuracy mostly matches with top-5 accuracy. The highest absolute top-5 accuracy gain is observed for encoder-decoder cross-modal Text+Chem T5 models. For an inversed evaluation under the AMORE framework (i.e., to retrieve an original SMILES x_i string given an augmented one x'_i), please see Appx. B. This performance is generally similar to the one presented in Table 1.

4.2 AMORE and downstream tasks

While our experiments on augmented molecule retrieval explored the robustness of an embedding space of chemical LM to SMILES augmentations, another interesting unexplored question is the robustness of these models to augmentations when solving a downstream task. We utilize two benchmarks: 1) MoleculeNet (Wu et al., 2018b) with chemistry tasks and 2) the ChEBI-20 dataset for the molecule captioning task. The MoleculeNet benchmark is the established standard in the research community to assess and compare the performance of models on various molecular property prediction tasks, spanning topics from quantum mechanics to physiology. We consider 9 tasks from it: three regression tasks (Lipophilicity, ESOL, FreeSolv), 3 binary classification tasks (HIV, BBBP, BBPA), and 3 multilabel classification tasks (Tox21, ToxCast, SIDER). Part of the results are presented in Figure 2, and full results are provided in Appendix J.

Augmented SMILES lead to degraded performance on chemical tasks Experiments showed metrics generally degrade on augmented MoleculeNet test sets (Fig. 2). For example, RMSE on the ESOL regression task increased from 0.87 to 7.93 with hydrogen addition. However, not all augmentations had the same impact, with cyclic augmentations having a smaller effect (0.93 to 0.99 for Text+ChemT5-standard). The impact of augmentations was more distinct in binary classification (BACE). PubChemDeBerta accuracy dropped from 0.8 to 0.38 with hydrogen addition, with intermediate drops for other augmentations. The major part of the model's accuracy range in binary classification (BBBP) is the following: original—cycle—Kekule—canon—hydrogen.

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

Table 3: Detailed evaluation results of ChemLMs for the ChEBI-20 test set: top-1 accuracy (Acc@1, %) for matching of distributed representations of molecules with their augmentations and ROUGE2 and METEOR for matching of textual outputs of LMs with gold descriptions (molecule captioning task). Here, canon refers to RDKit canonicalization, hydro to Hydrogen explicit addition, kekul to Kekulization, and cycles to cycle renumbering.

For multilabel classification, BERT-based models (PubChemDeBerta, ChemBerta) outperformed T5-based models, suggesting the latter may not be well-suited for tasks with many classes. Additionally, we ranked all models for each augmentation type separately and found out that the rankings on augmented test sets are consistent with ranking on non-augmented test sets in general (see Appx. D).

Even simple augmentations are challenging Our results on augmented molecule retrieval (see Table 1 and Table 2) showed that cycle renumbering changes the representation space less than other augmentations. But on chemical tasks, metrics degrade even on these types of augmentations. It seems that even the slightest distribution shift in input SMILES can hinder the ability of a chemical LM to solve downstream tasks.

Captioning quality is consistent with AMORE From Table 3, the most significant drop in ROUGE and METEOR is observed for the *hydrogen addition* augmentation, which is consistent with our proposed AMORE metric. While ROUGE and METEOR require labor-intensive labeled datasets for evaluation, our proposed embedding distance-based AMORE framework supports zero-shot evaluation and only requires a set of SMILES strings. Though the correlation between Acc@1 and ROUGE/METEOR is not 100% (40%/40% in the case of canonicalization and 60%/32% in the case of kekulization), we still suppose that such behavior is partly caused by the fact that ROUGE/METEOR metrics are not ideal and may decrease for linguistically richer models. Thus, both datasets demonstrate that part of augmented molecule embeddings is the closest to its original ones. Still, it strongly changes from one augmentation to another, and among the tested models, there was not one of the best metrics on all datasets and augmentations.

Representation robustness correlates across different augmentations The flexibility of our framework allows us to take hidden representations from an arbitrary intermediate layer of a ChemLM. We explored how the retrieval-based top-1 accuracy changes over different Transformer layers. Fig. 3 presents the layer-wise AMORE metric for T5 models. An interesting finding is that layer-wise retrieval quality strongly correlates across varying augmentation types. For instance, Text+Chem T5-standard faces a significant top-1 accuracy drop on the 12th decoder layer for three of four augmentation types simultaneously. The same stands for SciFive's decoder. For MolT5's encoder and decoder, a notable performance drop is observed for the 3rd layer. However, layer dynamics is not consistent across different Chem LMs.

Levenshtein: discrepancy between different types of augmentations To further discover the root of ChemLM's performance degradation on augmented test sets, we explored the dependency between molecule captioning quality on CheBI20 and simpler SMILES string properties. In particular, for each augmented test set, we measure average string length and the Levenshtein distance between the original SMILES and an augmented one. For each pair of original and augmented SMILES, we define Levenshtein ratio as the ratio between their Levenshtein distance and the length of the original SMILES string. Additionally, we include the Spearman's correlations between the target metrics, such as ROUGE1 and METEOR,

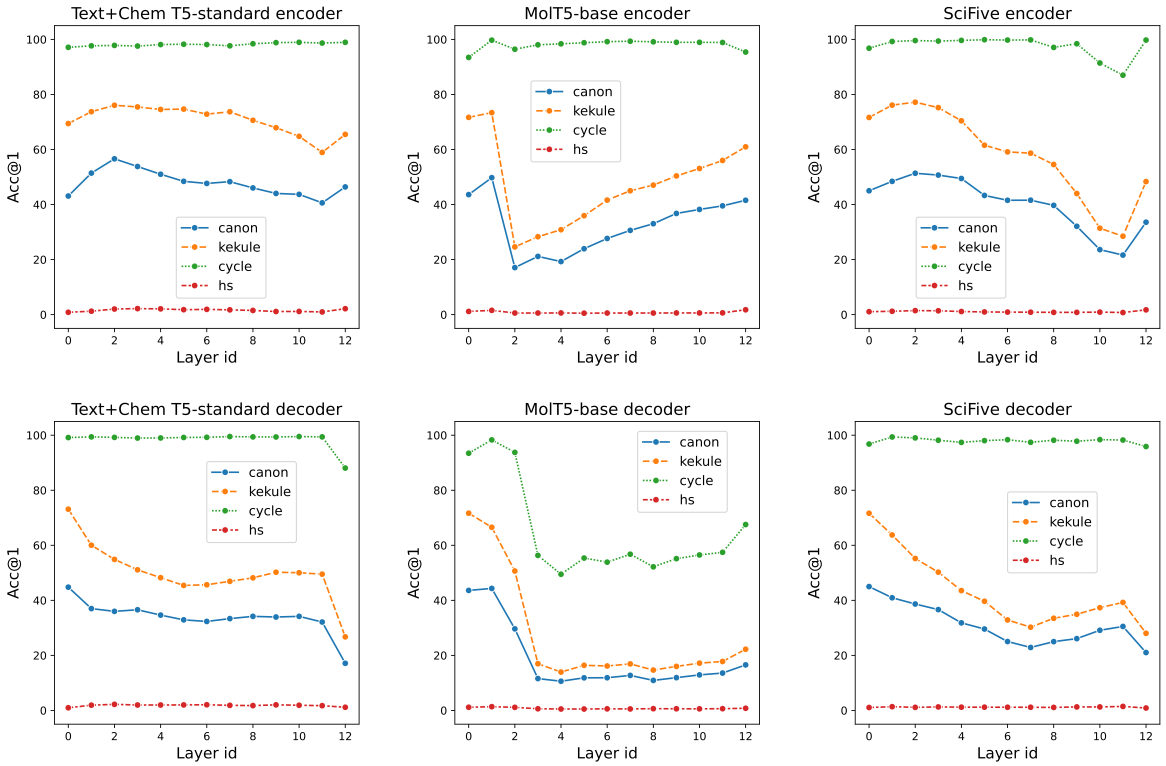

Figure 3: Top-1 retrieval accuracy (Acc@1) on CheBI-20 dataset calculated for hidden representations for different layers of encoder-decoder chemical LMs. The 0-th layer is the initial token embeddings (embedding layer) before any Transformer layers. The first row presents the results for encoders; the second row stands for decoders.

and the Levenshtein ratio for MolT5 model. The results are shown in the Table 11 of the Appendix I. While high string length (Levenshtein ratio for hydrogen augmentation is three times larger than for canonicalization or kekulization case) could partially explain poor generalization on hydrogen addition augmentation, low correlation values between the target metrics and Levenshtein ratio indicates that string variation is not the only challenge. A deeper insight into generalization limitations on augmented data requires a future work.

5 Discussion

In this paper, we release a general framework for analyzing knowledge awareness of modern LMs in the chemical domain. While we rely on L2 distance as a similarity distance throughout all our experiments, an arbitrary embedding similarity measure can be employed. Similarly, possible augmentation types are not limited to the ones considered in our research and can be extended. This flexibility might open new avenues for interpretation and analysis for LMs in the chemical domain.

Our experiments have shed light on the research question formulated in the Introduction and revealed a few critical limitations of the existing LMs in chemistry-related tasks. First, the embedding space of chemical LMs is not robust even to simple augmentations of SMILES strings known as identity transformations of molecules in chemistry. While robustness to these augmentations can vary across different model layers, no intermediate layer would be stable to SMILES augmentations. Second, the performance of chemical LMs at downstream tasks, such as molecule captioning, can be significantly limited when being passed an out-of-distribution (OOD) input. These two findings demonstrate that **the existing chemical LMs have problems with distinguishing the same molecules in different representations** during NLP-inspired pre-training procedures. They overfit on a specific format of input molecular string representations rather than truly gain an understanding of molecules. Finally, cross-modal chemical LMs tend to be more robust to OOD input samples, highlighting the importance of further developing multimodal models for chemistry and NLP. Meanwhile, metrics for the isomers dataset are lower and show minimal differences across models, likely attributed to the dataset's structure comprising isomeric aromatic compounds with identical molecular formulas and atom counts.

The key idea is that chemical models must accurately translate augmented SMILES into molecular structures. Without fully understanding SMILES syntax and distinguishing same-structure SMILES, ChemLMs remain vulnerable to real-world data perturbations. This analysis aims to inform revisions to the established pipeline for learning chemical representations from NLP.

6 Conclusion

In this paper, we introduce AMORE, a novel method (Fig. 1) based on embedding distance and SMILES augmentation to explore and evaluate the trustworthiness of language model's representations of a chemical substance and its ability to recognize molecule structures in SMILES string representations. By using this method, we assessed the most popular chemical LMs for several benchmarks (ChEBI-20, QM9, and MoleculeNet). We propose to use an isomeric subset of the QM9 dataset, which is novel to this task.

Though the first attempts to study the impact of chemical augmentations on Text+Chem T5 and MolT5 for molecule captioning exist, this is limited to cross-domain generative architectures requiring NLP tokens, constraining the number of suitable models for evaluation. The key novelty of our paper lies in the proposed probing scheme. It is the first application of computation of distances between embeddings for benchmarking chemical LLMs. As a result, our AMORE framework drastically extends this scope for evaluating and comparing models in domain-specific diverse architectures, including encoder-only versus generative models, as well as uni-modal LMs (with molecule atom tokens only) versus cross-modal models (atom + NLP tokens). It is important to emphasize that our method exploits unique specifics of the chemical domain. In contrast with typical NLP tasks, our augmentations lead to the creation of total synonyms of a molecule, which are absent in general words of natural language. Our framework opens avenues for future research, ranging from understanding the functionality of molecule SMILES representations in LMs to addressing weaknesses in chemical tasks and enhancing efficiency.

Limitations

First, we evaluated modes that are publicly available at HuggingFace (HF) (links in Appendix E). We note that there are other popular models such as Chemformer (`https://github.com/MolecularAI/Chemformer`), Molformer (`https://github.com/IBM/molformer`) and T5Chem (`https://github.com/HelloJocelynLu/t5chem`), which we failed to plug as HF checkpoints. Second, the evaluated models primarily focus on the sequence format of molecules, but it is important to consider in future other formats, such as 3D structures, which also hold significant importance. Third, we emphasize that the evaluated models were developed for research purposes and may contain unintended biases, and any molecules generated by them should undergo thorough evaluation through standard clinical testing. Furthermore, SELFIES (Krenn et al., 2022) and other molecule naming systems are also widespread in the chemical field. In our research we have focused on SMILES due to its popularity, but the augmentations on other systems are yet to be explored.

Ethics Statement

The models and datasets used in this work are publicly available for research purposes. The incorporation of AI into applied chemistry brings forth a variety of risks and ethical dilemmas. First, the direct implementation of AI-generated predictions, potentially hazardous or dangerous, without rigorous validation could result in human injuries, casualties, and damage to laboratory facilities. Second, the absence of proper oversight could lead to the misuse of chemical language models and AI in general, potentially facilitating the production of dangerous and illegal chemical compounds, with significant ethical and societal consequences. To address these concerns, it is essential to develop and implement safe ethical guidelines for the development and deployment of AI in chemistry.

Acknowledgments

This work was supported by a grant for research centers in the field of artificial intelligence, provided by the Analytical Center for the Government of the Russian Federation in accordance with the subsidy agreement (agreement identifier 000000D730321P5Q0002) and the agreement with the Ivannikov Institute for System Programming of the Russian Academy of Sciences dated November 2, 2021 No. 70-2021-00142.

We acknowledge the computational resources of HPC facilities at the HSE University.

References

Mark Aizerman and Fuad Aleskerov. 1995. Theory of choice. vol. 38. *Studies in Mathematical and Managerial Economics. North-Holland*, page 136.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

A Patrícia Bento, Anne Hersey, Eloy Félix, Greg Landrum, Anna Gaulton, Francis Atkinson, Louisa J Bellis, Marleen De Veij, and Andrew R Leach. 2020. An open source chemical structure curation pipeline using RDKit. *Journal of Cheminformatics*, 12:1–16.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Gayane Chilingaryan, Hovhannes Tamoyan, Ani Tevosyan, Nelly Babayan, Lusine Khondkaryan, Karen Hambardzumyan, Zaven Navoyan, Hrant Khachatrian, and Armen Aghajanyan. 2022. Bartsmiles: Generative masked language models for molecular representations. *arXiv preprint arXiv:2211.16349*.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. 2020. Chemberta: large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*.

Noam Chomsky. 1957. *Syntactic structures*. The Hague: Mouton.

Dimitrios Christofidellis, Giorgio Giannone, Jannis Born, Ole Winther, Teodoro Laino, and Matteo Manica. 2023. Unifying molecular and textual representations via multi-task language modelling. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 6140–6157. PMLR.

John S. Delaney. 2004. ESOL: estimating aqueous solubility directly from molecular structure. *J. Chem. Inf. Model.*, 44(3):1000–1005.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. 2022a. Translation between molecules and natural language. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 375–413, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. 2022b. Translation between molecules and natural language. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 375–413, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Carl Edwards, ChengXiang Zhai, and Heng Ji. 2021. Text2Mol: Cross-modal molecule retrieval with natural language queries. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 595–607.

Veronika Ganeeva, Kuzma Khrabrov, Artur Kadurin, Andrey Savchenko, and Elena Tutubalina. 2024. Chemical language models have problems with chemistry: A case study on molecule captioning task. In *The Second Tiny Papers Track at ICLR 2024*.

Anna Gaulton, Louisa J. Bellis, A. Patrícia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, and John P. Overington. 2012. Chembl: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.*, 40(Database-Issue):1100–1107.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. 2020. Message passing neural networks. *Machine learning meets quantum physics*, pages 199–214.

Landrum Greg et al. 2022. RDKit: open-source cheminformatics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

John J. Irwin, Khanh G. Tang, Jennifer Young, Chinzorig Dandarchuluun, Benjamin R. Wong, Munkhzul Khurelbaatar, Yurii S. Moroz, John W. Mayfield, and Roger A. Sayle. 2020. ZINC20 - A free ultralarge-scale chemical database for ligand discovery. *J. Chem. Inf. Model.*, 60(12):6065–6073.

Ross Irwin, Spyridon Dimitriadis, Jiazhen He, and Esben Jannik Bjerrum. 2022. Chemformer: a pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology*, 3(1):015022.

Jeff Johnson, Matthijs Douze, and Herve Jegou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(03):535–547.

Karl. 2024. Gpt2 zinc 87m.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A. Shoemaker, Paul A. Thiessen, Bo Yu, Leonid Zaslavsky, Jian Zhang, and Evan E. Bolton. 2023. Pubchem 2023 update. *Nucleic Acids Res.*, 51(D1):1373–1380.

Sunghwan Kim, Paul A Thiessen, Evan E Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A Shoemaker, et al. 2016. Pubchem substance and compound databases. *Nucleic acids research*, 44(D1):D1202–D1213.

PS Kostenetskiy, RA Chulkevich, and VI Kozyrev. 2021. Hpc resources of the higher school of economics. In *Journal of Physics: Conference Series*, volume 1740, page 012050. IOP Publishing.

Mario Krenn, Qianxiang Ai, Senja Barthel, Nessa Carson, Angelo Frei, Nathan C. Frey, Pascal Friederich, Théophile Gaudin, Alberto Alexander Gayle, Kevin Maik Jablonka, Rafael F. Lameiro, Dominik Lemm, Alston Lo, Seyed Mohamad Moosavi, José Manuel Nápoles-Duarte, AkshatKumar Nigam, Robert Pollice, Kohulan Rajan, Ulrich Schatzschneider, Philippe Schwaller, Marta Skreta, Berend Smit, Felix Strieth-Kalthoff, Chong Sun, Gary Tom, Guido Falk von Rudorff, Andrew Wang, Andrew D. White, Adamo Young, Rose Yu, and Alán Aspuru-Guzik. 2022. Selfies and the future of molecular string representations. *Patterns*, 3(10):100588.

Michael Kuhn, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. 2016. The sider database of drugs and side effects. *Nucleic acids research*, 44(D1):D1075–D1079.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Mufei Li, Jinjing Zhou, Jiajing Hu, Wenxuan Fan, Yangkang Zhang, Yaxin Gu, and George Karypis. 2021. Dgl-lifesci: An open-source toolkit for deep learning on graphs in life science. *ACS Omega*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Daniel Lowe. 2017. Chemical reactions from US patents (1976-Sep2016).

Daniel Mark Lowe. 2012. *Extraction of chemical structures and reactions from the literature*. Ph.D. thesis, University of Cambridge.

Jieyu Lu and Yingkai Zhang. 2022. Unified deep learning model for multitask reaction predictions with explanation. *Journal of Chemical Information and Modeling*, 62(6):1376–1387.

Raimund Mannhold, Gennadiy I Poda, Claude Ostermann, and Igor V Tetko. 2009. Calculation of molecular lipophilicity: State-of-the-art and comparison of log p methods on more than 96,000 compounds. *Journal of pharmaceutical sciences*, 98(3):861–893.

Damian Marino, P.J. Peruzzo, and Andrey Toropov. 2001. Qsar carcinogenic study of polycyclic aromatic hydrocarbons based on topological descriptors derived from the detour matrix and correlation weights of local graph invariants. *Chemical Physics Letters - CHEM PHYS LETT*, 2001:111–126.

Ines Filipa Martins, Ana L. Teixeira, Luis Pinheiro, and André O. Falcão. 2012. A bayesian approach to *in Silico* blood-brain barrier penetration modeling. *J. Chem. Inf. Model.*, 52(6):1686–1697.

David L. Mobley and J. Peter Guthrie. 2014. Freesolv: a database of experimental and calculated hydration free energies, with input files. *J. Comput. Aided Mol. Des.*, 28(7):711–720.

Long N Phan, James T Anibal, Hieu Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltekian, and Grégoire Altan-Bonnet. 2021. Scifive: a text-to-text transformer model for biomedical literature. *arXiv preprint arXiv:2106.03598*.

Dragomir R Radev, Hong Qi, Harris Wu, and Weiguo Fan. 2002. Evaluating web-based question answering systems. In *LREC*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole Von Lilienfeld. 2014. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data*, 1(1):1–7.

Ann M Richard, Ruili Huang, Suramya Waidyanatha, Paul Shinn, Bradley J Collins, Inthirany Thillainadarajah, Christopher M Grulke, Antony J Williams, Ryan R Lougee, Richard S Judson, et al. 2020. The tox21 10k compound library: collaborative chemistry advancing toxicology. *Chemical Research in Toxicology*, 34(2):189–216.

Mark Rofin, Vladislav Mikhailov, Mikhail Florinsky, Andrey Kravchenko, Tatiana Shavrina, Elena Tutubalina, Daniel Karabekyan, and Ekaterina Artemova. 2023. Vote'n'rank: Revision of benchmarking with social choice theory. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 670–686, Dubrovnik, Croatia. Association for Computational Linguistics.

Lars Ruddigkeit, Ruud Van Deursen, Lorenz C Blum, and Jean-Louis Reymond. 2012. Enumeration of 166 billion organic small molecules in the chemical universe database gdb-17. *Journal of chemical information and modeling*, 52(11):2864–2875.

Andrey Sakhovskiy, Natalia Semenova, Artur Kadurin, and Elena Tutubalina. 2023. Graph-enriched biomedical entity representation transformer. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 14th International Conference of the CLEF Association, CLEF 2023, Thessaloniki, Greece, September 18-21, 2023, Proceedings*, volume 14163 of *Lecture Notes in Computer Science*, pages 109–120. Springer.

Andrey Sakhovskiy, Natalia Semenova, Artur Kadurin, and Elena Tutubalina. 2024. Biomedical entity representation with graph-augmented multi-objective transformer. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4626–4643, Mexico City, Mexico. Association for Computational Linguistics.

Maximilian G Schuh, Davide Boldini, and Stephan A Sieber. 2024. Twinbooster: Synergising large language models with barlow twins and gradient boosting for enhanced molecular property prediction. *arXiv preprint arXiv:2401.04478*.

Teague Sterling and John J. Irwin. 2015. Zinc 15 – ligand discovery for everyone. *Journal of Chemical Information and Modeling*, 55(11):2324–2337. PMID: 26479676.

Govindan Subramanian, Bharath Ramsundar, Vijay Pande, and Rajiah Aldrin Denny. 2016. Computational modeling of β-secretase 1 (bace-1) inhibitors using ligand based approaches. *Journal of chemical information and modeling*, 56(10):1936–1949.

Elena Tutubalina, Artur Kadurin, and Zulfat Miftahutdinov. 2020. Fair evaluation in concept normalization: a large-scale comparative analysis for BERT-based models. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6710–6716, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

David Weininger. 1988. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, 28:31–36.

Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. 2018a. MoleculeNet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530.

Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. 2018b. Moleculenet: a benchmark for molecular machine learning. *Chem. Sci.*, 9:513–530.

Xiao-Chen Zhang, Cheng-Kun Wu, Jia-Cai Yi, Xiang-Xiang Zeng, Can-Qun Yang, Ai-Ping Lu, Ting-Jun Hou, and Dong-Sheng Cao. 2022. Pushing the boundaries of molecular property prediction for drug discovery with multitask learning bert enhanced by smiles enumeration. *Research*, 2022:0004.

A Augmentation examples

Examples of molecules and their augmented versions are provided in Table 4.

B Augmented-original AMORE

Table 5 presents the Top-1 and Top-5 retrieval accuracy for the inversed version of AMORE: for retrieval of original non-augmented X SMILES given augmented ones X'. Overall, the results are comparable to the ones in Table 1, which indicates that these two tasks are of similar complexity for chemical LMs under consideration.

C Molecule captioning evaluation

For evaluation on molecule captioning, we utilized the luna-nlg framework, which is available at `https://pypi.org/project/luna-nlg/` and RDKit (`https://www.rdkit.org`). Trained models, data, and our source code will be published upon acceptance.

TYPE	RESULT
original	`CNCCC(C1=CC=CS1)OC2=CC=CC3=CC=CC=C32.Cl`
canon	`CNCCC(Oc1cccc2ccccc12)c1cccs1.Cl`
hydro	`[CH3][NH][CH2][CH2][CH]([O][c]1[cH][cH][cH][c]2[cH][cH][cH][cH][c]12)[c]1[cH][cH][cH][s]1.[ClH]`
kekul	`CNCCC(OC1=CC=CC2=CC=CC=C12)C1=CC=CS1.Cl`
cycle	`CNCCC(C4=CC=CS4)OC2=CC=CC7=CC=CC=C72.Cl`

Table 4: Example of molecular SMILES from a dataset and after our transformations. Here, canon refers to RDKit canonicalization, hydro to Hydrogen explicit addition, kekul to Kekulization, and cycle to cycle renumbering

Model	**Canon**		**Hydro**		**Kekul**		**Cycle**	
	@1	@5	@1	@5	@1	@5	@1	@5
Generative models								
Text+Chem T5-standard	52.52	69.18	2.88	6.39	71.09	86.97	99.09	99.67
Text+Chem T5-augm	50.58	67.12	3.88	8.97	71.3	86.42	98.94	99.97
MolT5-base	41.55	57.79	1.76	3.48	61.03	79.42	95.39	98.64
MolT5-large	50.18	66.52	1.52	3.03	57.42	72.73	99.73	100
BARTSmiles	25.18	36.55	0.79	1.45	36.36	51.76	73.42	82.15
ZINC-GPT	25.03	34.12	0.7	0.94	36.73	47.94	82.03	89.21
SciFive	33.58	50.18	1.73	3.55	48.33	66.91	99.76	100
Encoder-only models								
PubChemDeBERTa	35.12	53.15	1.3	2.36	53.73	72.91	99.76	100
ChemBERT-ChEMBL	25.09	37.61	1.03	2.06	34.42	51.24	85.61	91.21
ChemBERTa	26.61	25.79	40.42	0.88	1.82	43.15	65.33	96.97
ZINC-RoBERTa	24.64	36.36	1.03	1.76	34.33	49.73	95.09	99.18
Graph Neural Network model								
GNN	91.96	97.88	91.96	97.88	91.96	97.88	100.	100.

Table 5: Top-1 and Top-5 accuracy (@1/@5) of ChemLMs and GNN for matching of distributed representations of augmented SMILES strings to the original (non-augmented) ones on CheBI-20 dataset.

D Chemical LM Ranking

To explore how the ranking of ChemLMs on augmented test sets changes compared to non-augmented data, we conduct our experiments on nine datasets MoleculeNet datasets as follows. Each model is trained on the original train set provided in MoleculeNet and evaluated on both the original test set and four augmented test sets. Next, we rank all models with respect to their performance on a given test set type (either an original one or one of four augmented ones) using the Vote'n'Rank framework (Rofin et al., 2023). The framework is designed for ranking systems in multi-task benchmarks under the principles of the social choice theory (Aizerman and Aleskerov, 1995). We follow recommendations from (Rofin et al., 2023) and use Copeland rule to select the system that beats all the others in pairwise comparison. Copeland chooses the system that dominates the others in more cases and is dominated by the least.

The results are presented in Table 6. Overall, all augmentations except for hydrogen addition do not seem to shuffle the original ranking too much. For instance, Zinc-RoBERTa and PubChemDeBERTa achieve rank 1 and 2, respectively, on four of five test sets. Similarly, MolT5-base placed last on all augmentations except for hydrogen addition. It seems that encoder-decoder architectures are more stable to hydrogen addition on downstream tasks than encoder-only architectures as 4 of top 5 places are achieved by MolT5-large, MolT5-base, Text+Chem T5-augm, and SciFive.

E Dataset and Model Details

The MoleculeNet benchmark (Wu et al., 2018b) consists of 17 different chemical tasks. In this work, we consider 9 tasks from it: three regression tasks (Lipophilicity, ESOL, FreeSolv), 3 binary classification tasks (HIV, BBBP, BBPA), and 3 multilabel

Rank	**Test set**				
	Original	**Canon**	**Hydro**	**Kekul**	**Cycles**
1	‡	‡	◇	‡	‡
2	†	†	♥	†	†
3	□	♦	♠	♦	□
4	♠	□	†	□	♠
5	♦	♠	△	♠	♦
6	♣	◇	♣	♣	♣
7	△	△	‡	◇	△
8	◇	♣	□	△	◇
9	♥	♥	♦	♥	♥

Table 6: ChemLM rankings with respect to Vote'n'Rank framework's Copeland score calculated on 9 downstream tasks from the MoleculeNet benchmark for different augmentation types. Here, canon refers to RDKit canonicalization, hydro to Hydrogen explicit addition, kekul to Kekulization, and cycles to cycle renumbering.
Models: ‡=ZINC-RoBERTa, †=PubChemDeBERTa, □=ChemBerta, ♣=Text+Chem T5-augm, ♦=Text+Chem T5-standard, ♣=Text+Chem T5-augm, △=SciFive, ◇=MolT5-large, ♥=MolT5-base.

classification tasks (Tox21, ToxCast, SIDER). For our experiments, we employ the MoleculeNet's 80/10/10% splits into train, validation, and test sets. The MoleculeNet benchmark is available at `https://moleculenet.org`.

We trained each model on each MoleculeNet task for 50 epochs with learning rates of $1 \cdot 10^{-5}$ and $1 \cdot 10^{-3}$ for encoder-only and encoder-decoder architectures, respectively. For prediction, we loaded model weights from the best epoch in terms of target metric on the validation set: accuracy for classification tasks and RMSE for regression tasks. Encoder-Decoder T5 models are trained for text generation given prompts listed in Table 7 as input.

BBBP (Martins et al., 2012) is a task to classify if a molecule penetrates the Blood-Brain barrier. The corresponding dataset consists of 2053 compounds.

HIV (Wu et al., 2018b) is a task to predict if a molecule can inhibit HIV replication. The corresponding dataset consists of 4000 compounds. Screening results were evaluated and placed into three categories: confirmed inactive (CI), confirmed active (CA), and confirmed moderately active (CM), and further combined, making it a classification task between inactive (CI) and active (CA and CM).

BACE (Subramanian et al., 2016) is a task to predict if a molecule will be an inhibitor of human beta-secretase (BACE-1). The corresponding dataset consists of 1513 compounds.

Tox21 (Richard et al., 2020) is the task of predicting qualitative toxicity measurements on 12 biological targets, including nuclear receptors and stress response pathways. The corresponding dataset consists of 7831 compounds.

ClinTox (Wu et al., 2018b) dataset compares drugs approved by the FDA and drugs that have failed clinical trials for toxicity reasons. The dataset includes two classification tasks for 1491 drug compounds with known chemical structures: (1) clinical trial toxicity (or absence of toxicity) and (2) FDA approval status.

SIDER (Kuhn et al., 2016) is a task to classify marketed drugs and adverse drug reactions (ADR) into 27 system organ classes. The corresponding dataset consists of 1427 compounds.

ESOL (Delaney, 2004) is a task to predict the water solubility of a compound. The corresponding dataset consists of 1128 compounds.

FreeSolv (Mobley and Guthrie, 2014) is a task to predict the hydration-free energy of small molecules in water. The calculated values are derived from alchemical free energy calculations using molecular dynamics simulations. The corresponding dataset consists of 642 compounds.

Lipophilicity (Mannhold et al., 2009) is a task to predict the octanol/water distribution coefficient ($logD$ at pH 7.4). The corresponding dataset consists of 1128 compounds derived from the ChEMBL database. This property is an important feature of drug molecules affecting membrane permeability and solubility.

The ChEBI-20 dataset used for experiments is available at `https://github.com/blender-nlp/MolT5/tree/main/ChEBI-20_data`. The QM9 dataset is available at `http://quantum-machine.org/datasets/`.

The GNN checkpoint is available at `https://github.com/awslabs/dgl-lifesci/tree/master/examples/molecule_embeddings`.

Table 8 lists the Hugging Face checkpoints for models used in this study. Table 9 summarizes the parameter count and domain of each utilized LM.

Text+Chem T5 (Christofidellis et al., 2023) is a multi-task, cross-domain language model that unifies natural language and chemical representations. It employs a shared T5 (Raffel et al., 2020) encoder-decoder to learn from aligned text-SMILES pairs. For our experiments, we adopted two Text+Chem T5 *base*-sized models: (i) *Text+Chem T5-standard*, which is pre-trained on these 11.5M samples, and

Task	Dataset	T5 prompt	# Samples
Regression	ESOL	What is water solubility of SMILES?	1,128
	FreeSolv	What is hydration free energy of SMILES in water?	642
	Lipophilicity	How lipophilic is SMILES?	4,200
Binary classification	BACE	Please evaluate the ability of SMILES to inhibit human beta-secretase.	1,513
	BBBP	Can SMILES penetrate the BBB?	2,039
	HIV	Is SMILES an HIV inhibitor?	41,127
Multilabel classification	ClinTox	Given drug compound SMILES predict its toxicity and FDA approval status.	1,478
	SIDER	Given drug compound SMILES, predict the organ classes for which it causes adverse reactions.	1,427
	Tox21	Given molecule SMILES, predict its toxicity measurements.	7,831

Table 7: Overview of the adopted MoleculeNet datasets. For each dataset, a prompt for fine-tuning a sequence-to-sequence T5 model and samples count are provided.

(ii) *Text+Chem T5-augm* which is pre-trained on an augmented version of this corpus that consists of 33.5M paired samples.

MolT5 (Edwards et al., 2022b) is a self-supervised learning framework for jointly training a model on molecule captioning and text-based molecule generation tasks. The model employs a multi-task pre-training pipeline (Raffel et al., 2020) to learn from 100M SMILES strings from the ZINC-15 database (Sterling and Irwin, 2015) and natural language texts from the C4 (Raffel et al., 2020) corpus.

PubChemDeBERTa (Schuh et al., 2024) adopts DeBERTa V3 (He et al., 2023) encoder to learn molecular representations on PubChem (Kim et al., 2023) via the replaced token detection pre-training task. The model simultaneously adopts a Siamese neural network architecture to learn from biological assays, molecular fingerprints, and textual features (such as a molecule's description and title). The authors released two versions of the pre-trained model: (i) a base one (ii) and an augmented one, which was trained on augmented textual descriptions. In our work, we experimented with the augmented version as it achieved higher perplexity on a test set (Schuh et al., 2024).

ChemBERT-ChEMBL is a BERT-based (Devlin et al., 2019) model pre-trained on 1.7M molecules in SMILES format from the ChemBL (Gaulton et al., 2012) database via the masked-language modeling (MLM) objective.

ChemBERTa (Chithrananda et al., 2020) is a RoBERTa-based (Liu et al., 2019) molecular representation model which is pre-trained on 100K SMILES strings from the ZINC (Sterling and Irwin, 2015) benchmark via the MLM objective.

BARTSmiles (Chilingaryan et al., 2022) is a BART-like (Lewis et al., 2020) sequence-to-sequence molecular representation model pre-trained on 1.7B SMILES samples from the Zinc20 (Irwin et al., 2020) chemical database.

ZINC-GPT is a GPT-like (Radford et al., 2019) autoregressive language model trained on 480K SMILES strings from the ZINC (Sterling and Irwin, 2015) database.

ZINC-RoBERTa is a RoBERTa-based (Liu et al., 2019) molecular representation model which is pre-trained on 480K SMILES strings from the ZINC (Sterling and Irwin, 2015) database via the MLM objective.

SciFive is a uni-modal textual T5-based model pre-trained on the union of general-domain C4 corpus and 32M abstracts from the PubMed database[2]. We adopt the model for our experiments to investigate if special chemical LMs are needed or if simple training of a universal LM with both textual and chemical modalities is enough for chemistry-related tasks.

All the experiments in this paper were conducted on a single machine with Nvidia V100 GPU and a 8-core CPU (Kostenetskiy et al., 2021).

[2] https://pubmed.ncbi.nlm.nih.gov

Model	**HuggingFace checkpoint**
Text+Chem T5-standard (Christofidellis et al., 2023)	GT4SD/multitask-text-and-chemistry-t5-base-standard
Text+Chem T5-augm (Christofidellis et al., 2023)	GT4SD/multitask-text-and-chemistry-t5-base-augm
MolT5-base (Edwards et al., 2022b)	laituan245/molt5-base-smiles2caption
MolT5-large (Edwards et al., 2022b)	laituan245/molt5-large-smiles2caption
SciFive (Phan et al., 2021)	razent/SciFive-base-Pubmed
PubChemDeBERTa (Schuh et al., 2024)	mschuh/PubChemDeBERTa-augmented
ChemBERT-ChEMBL (Zhang et al., 2022)	jonghyunlee/ChemBERT_ChEMBL_pretrained
ChemBERTa (Chithrananda et al., 2020)	seyonec/ChemBERTa-zinc-base-v1
BARTSmiles (Chilingaryan et al., 2022)	gayane/BARTSmiles
ZINC-GPT	entropy/gpt2_zinc_87m
ZINC-RoBERTa	entropy/roberta_zinc_480m

Table 8: HuggingFace checkpoints used in our expriments.

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

Table 9: Domain and parameter count for models used in this study. “Chem” and “Text” are uni-modal chemical and textual models. “Cross” stands for cross-domain (bi-modal) language and chemistry LMs.

F Qualitative Analysis

An example of the outputs of the models (Text+Chem T5-augm predictions for duloxetine hydrochloride) is shown in Table 10.

All in all, **none of the predicted substances and reactions are correct**.

Cycle augmentation: the model has correctly predicted the role of duloxetine hydrochloride (as an antidepressant and a serotonin uptake inhibitor) and a precursor reaction with hydrochloric acid with a correct molar ratio. However, the model wrongly suggests Irinotecan as a precursor, and with this reaction, the presence of atomic sulfur in the resulting hydrochloride is not explained. Original test: the model suggests that duloxetine hydrochloride is a result of a reaction between 1-[2-(1-benzothiophen-3-yl)ethoxy]-4-(methylamino)butanal and hydrochloric acid. The medical role and substance class (hydrochloride) are correctly predicted. But 1-[2-(1-benzothiophen-3-yl)ethoxy]-4-(methylamino)butanal seems non-existent (it is not described on PubChem and not found in any chemical paper). However, molecules with components such as (-4-(methylamino)butanal and [2-(1-benzothiophen-3-yl)ethoxy]) exist in the training set. Canonical test: the model has correctly predicted the role of duloxetine hydrochloride (as an antidepressant and a serotonin uptake inhibitor) and a precursor reaction with hydrochloric acid with a correct molar ratio. However, the model suggests cis-diltiazem as a precursor, but this substance is a calcium channel blocker and an angina treatment drug. If this reaction was correct and possible, then this antidepressant would be produced in the stomach from the reaction of gastric acid and diltiazem. It is a problem because these predictions are fully incorrect and potentially dangerous in practice. Hydrogen augmentation: the model has given the worst predictions on this augmentation, as all components have been suggested incorrectly, the role of the substance is not predicted, and both precursor substances are also incorrect. Kekule: the model has correctly predicted the role of duloxetine hydrochloride (as an antidepressant and a serotonin uptake inhibitor) and a precursor reaction with hydrochloric acid with a correct molar ratio. However, the model suggests N-methyltryptamine as a precursor, which

Probing Test	Input	Output

original	CNCCC(C1=CC=CS1)
	OC2=CC=CC3=CC=CC=C32.Cl
		The molecule is a **hydrochloride salt** resulting from the reaction of **1-[2-(1-benzothiophen-3-yl)ethoxy]-4-(methylamino)butanal** with **1 mol eq.** of **hydrogen chloride**. It has a role as an **antidepressant** and a **serotonin uptake inhibitor**. It contains a **1-[2-(1-benzothiophen-3-yl)ethoxy]-4-(methylamino)butanal**.

canonical	CNCCC(Oc1cccc2ccccc12) c1cccs1.Cl	The molecule is a **hydrochloride** obtained by combining **equimolar** amounts of **cis-diltiazem** and **hydrogen chloride**. It has a role as an **antidepressant** and a **serotonin uptake inhibitor**. It contains a **cis-diltiazem(1+)**.
hydrogen	[CH3][NH][CH2][CH2][CH] ([O][c]1[cH][cH][cH] [c]2[cH][cH][cH][cH] [c]12)[c]1[cH][cH] [cH][s]1.[ClH]	The molecule is an **ammonium salt** resulting from the reaction of **equimolar** amounts of **scyllo-inositol** and **ammonia**. It contains a **scyllo-inositol(1+)**. It derives from a **hydride of a scyllo-inositol**.
kekulization	CNCCC(OC1=CC =CC2=CC=CC=C12)C1=CC=CS1.Cl	The molecule is a **hydrochloride** obtained by combining **N-methyltryptamine** with **one molar** equivalent of **hydrochloric acid**. It has a role as an **antidepressant** and a **serotonin uptake inhibitor**. It contains a **N-methyltryptamine(1+)**.

cycles	CNCCC(C4=CC=CS4)
	OC2=CC=CC7=CC=CC=C72.Cl
		The molecule is a **hydrochloride** obtained by combining **irinotecan** with one molar equivalent of **hydrochloric acid**. It has a role as an **antidepressant** and a **serotonin uptake inhibitor**. It contains an **irinotecan(1+)**.

Pubchem	CNCCC(C1=CC=CS1)
	OC2=CC=CC3=CC=CC=C32.Cl
		The molecule is a **duloxetine hydrochloride** in which the duloxetine moiety has S configuration. It has a role as an **antidepressant**. It contains a **(S)-duloxetine**.

Table 10: Examples of SMILES and descriptions generated by Chem+Text T5-base. We highlighted in bold the most important components of the description: precursors, role, and others.

is incorrect. This molecule is not present in train data, but it does exist and is found in acacia leaves and human urea as a result of bodily metabolic reactions. To sum up, the model does not generate predictions correctly both on augmented and non-augmented tests. But **molecular descriptions on augmented data are even further from the target description**. The model is able to correctly predict the role of duloxetine hydrochloride as an antidepressant and a serotonin uptake inhibitor and its precursor reaction with hydrochloric acid with a correct molar ratio on original SMILES for all augmentations excluding hydrogen addition. All the model's predictions seem to be based on the following high-frequency formula found in the training data within descriptions of substances containing ".Cl" : "The molecule is a hydrochloride obtained by combining ... with one molar equivalent of hydrochloric acid. It has a role as ... It contains a ... (1+)."

G Augmentations: chemical view

These models are not trained with augmentations specially. However, the rules for describing

Augmentation type	SMILES length (mean/std)	Levenshtein ratio with the original string (mean/std)	Correlation between Levenshtein ratio and ROUGE1	Correlation between Levenshtein ratio and METEOR
no augmentation	78.96 (80.29)	0	-	-
canon	74.71 (78.06)	0.47 (0.22)	-0.33	-0.34
hydro	153.36 (134.42)	1.49 (0.54)	-0.05	-0.09
cycles	78.97 (80.29)	0.04 (0.04)	-0.34	-0.33
kekul	76.98 (78.18)	0.40 (0.20)	-0.24	-0.22

Table 11: Levenshtain ratio between different types of augmentations

molecules using this text format do not have a single generally accepted standard, and some variations are allowed in how to write a molecule in this format. Even though the models were not trained on the augmented data specifically, the models trained on very large datasets are expected to understand basic simple rules from chemistry. Our augmentations are more like small changes in the order of words that do not affect their meaning than a fundamentally different format or another language. We expect that for a model in the inner layers of which there is an idea of the molecule's structure, and not only of its textual representation, then small changes in the recording of the molecule, and not in itself, should not pose a difficulty for the model. The representations of our models about molecules and augmented ones should be similar to the representations of language models (for example, in the semantic representations task) about sentences with the same meaning but with different word order: "Language models are used for biomedical and chemical tasks" and "Language models are used for chemical and biomedical tasks". These papers (models) highlight that models are training on various modalities before the main training for the task of generating descriptions and "overcome chemistry domain shortcoming of data scarcity" (Edwards et al., 2022a). However, the main problem is the difference between overcoming data scarcity in the chemical domain by training models on large data and overcoming data scarcity in the chemical domain by training models on various data. Our experiments suggest that large language models in the chemistry domain are large and language, but not chemistry, because these models can not overcome the problem of generating different descriptions for the same molecule. Even though the models were not trained on the augmented data specifically, the models trained on very large datasets are expected to understand basic simple rules from chemistry. We wanted to pay attention to the problem that good chemical models should recreate the structure of a molecule from a textual representation to a greater extent than relying solely on a textual representation.

SMILES supports the general "hydrogen suppression" approach common in chemoinformatics (Marino et al., 2001). However, SMILES with explicitly added hydrogen are also valid and supported by most chemical tools. Despite the convenience of the "hydrogen suppression", this approach has several problems. Here are a few examples: The implementation adopts an implicit hydrogens scheme: the number of implicit hydrogens is calculated as the difference between the "target valence" of an atom and its bond order. However, some atoms may have different valence in different molecules, and if the default valence is wrong, the user should explicitly define hydrogens. A similar issue arises when the molecule becomes charged and the "target valence" differs from the real. Molecules are 3D objects whose chemical/biological properties depend on the chirality. In some cases, the relative position of the hydrogen atom defines the chirality of the molecule, in which case it has to be included explicitly in the SMILES. Due to these reasons (in these cases, added hydrogens are obligatory), we can expect that chemical LLMs understand SMILES with added hydrogens.

H Augmentations: linguistic view

Why is the skill of understanding chemistry related to understanding syntactic rules of SMILES? We suggest these tests to explore LLMs' ability to understand chemistry because we assume that understanding syntactic rules relates to understanding chemistry.

In comparsion to natural language, Noam Chomsky in "Syntactic Structures" (Chomsky, 1957) proposed the idea of grammar as a mechanism that generates all grammatically correct sentences of a language and at the same time does not generate incorrect ones. By giving sentences in the SMILES language to the LM, we can expect that if it understands their composition and syntactic rules well, it will easily distinguish the same sentences from different ones.

If we compare these three sentences:

- *green colorless ideas sleep furiously*
- *colorless green ideas sleep furiously*
- *red colorless ideas sleep furiously*

then we probably agree that the first two of them are intended to convey the same information and semantics (and do it grammatically), but the last one it clearly conveys another one, although the first and last sentence have the character-by-character sequences of the longest length. We can compare these semantics even considering that the sentences have no meaning.

Our augmented tests do the same comparsion with SMILES: SMILES language has ambiguity, the same molecule can be represented by different strings. If we compare these three SMILES strings:

- *CC(=O)C1CC2OC2C1*
- *CC(=O)C1CC2C(C1)O2*
- *CC(=O)CC1(CC1)C=O*

then the first two of them are intended to convey the same molecule, but the last one it clearly conveys another one. It is not about totally different syntactic systems of SMILES writing rules, it is mostly about ambiguity in SMILES system. We define these differences as different types of augmentations, but most of our augmentations based exactly on SMILES ambiguity.

We agree that hydrogen addition is rare and not so widely used in the field of chemistry in comparsion to the other augmentations, but we include this augmentation in our framework to represent all spectrum of SMILES variability from the simplest (cycles renumbering) to the most difficult.

I Levenshtein: discrepancy between different types of augmentations

We conjecture that the main source of the discrepancy between augmentations is the difference between an original SMILES string and its augmentation. As a partial confirmation of the conjecture, we list the properties of the following distributions on the ChEBi dataset: augmented and original SMILES length, the Levenshtein ratio (the ratio between the Levenshtein distance and the length of an original string) between original SMILES and its augmentation; additionally, we include the Spearman's correlations between the target metrics and the Levenshtein ratio for MolT5 model. This data is shown in the Table 11. For instance, the Levenshtein ratio between a SMILES string and its Hydrogens augmentation is approximately three times larger than for canonicalization or Kekulization cases.

J Full results on Chemical Tasks

Model	Orig	cycle	hydro	kekul	canon
ChemBerta	0.7697	0.7829	0.3750	0.6382	0.6711
PubChemDeBERTa	0.8026	0.7895	0.3816	0.6908	0.6776
ChemBERT-ChEMBL	0.7829	0.7763	0.6184	0.5789	0.5000
ZINC-RoBERTa	0.8158	0.7763	0.4474	0.6250	0.6382
Text+Chem_T5-standard	0.7829	0.7500	0.6316	0.6776	0.7303
Text+Chem_T5-augm	0.7632	0.7303	0.5263	0.6711	0.6711
MolT5-base	0.6184	0.6184	0.6184	0.6184	0.6184
MolT5-large	0.6184	0.6184	0.6184	0.6184	0.6184
SciFive	0.7105	0.7039	0.6579	0.6645	0.6711
GCN	0.8289	0.8200	0.8026	0.8289	0.8289

Table 12: Accuracy metrics of different models for the BACE task. Here, canon refers to RDKit canonicalization, hydro to Hydrogen explicit addition, kekul to Kekulization, and cycles to cycle renumbering.

Model	Orig	cycle	hydro	kekul	canon
ChemBerta	0.9655	0.9638	0.9562	0.9633	0.9655
PubChemDeBERTa	0.9652	0.9640	0.9616	0.9628	0.9655
ChemBERT-ChEMBL	0.9664	0.9667	0.9365	0.9572	0.9662
ZINC-RoBERTa	0.9667	0.9667	0.9616	0.9657	0.9669
Text+Chem_T5-standard	0.9628	0.9660	0.9482	0.9640	0.9628
Text+Chem_T5-augm	0.9589	0.9635	0.9550	0.9601	0.9589
MolT5-base	0.9616	0.9616	0.9616	0.9616	0.9616
MolT5-large	0.9616	0.9616	0.9589	0.9616	0.9616
SciFive	0.9628	0.9626	0.9582	0.9630	0.9628
GCN	0.9737	0.9711	0.9523	0.9737	0.9737

Table 13: Accuracy metrics of different models for the HIV task. Here, canon refers to RDKit canonicalization, hydro to Hydrogen explicit addition, kekul to Kekulization, and cycles to cycle renumbering.

Model	Orig	cycle	hydro	kekul	canon
ChemBerta	0.9072	0.8918	0.5103	0.7990	0.7474
PubChemDeBERTa	0.9072	0.8918	0.4227	0.7680	0.7010
ChemBERT-ChEMBL	0.8918	0.8866	0.6856	0.7990	0.7320
ZINC-RoBERTa	0.9021	0.8918	0.5052	0.7680	0.7062
Text+Chem_T5-standard	0.8814	0.8814	0.6649	0.7990	0.7784
Text+Chem_T5-augm	0.8918	0.8814	0.6392	0.7577	0.7474
MolT5-base	0.8608	0.8608	0.7010	0.7784	0.7784
MolT5-large	0.9021	0.8918	0.6392	0.7268	0.7784
SciFive	0.9021	0.8814	0.6289	0.7268	0.7680
GCN	0.8542	0.8542	0.7917	0.8281	0.8281

Table 14: Accuracy metrics of different models for the BBBP task. Here, canon refers to RDKit canonicalization, hydro to Hydrogen explicit addition, kekul to Kekulization, and cycles to cycle renumbering.

Model	Orig	Cycle	Hydro	Kekul	Canon
ChemBerta	0.9755	0.9755	0.9301	0.4755	0.9301
PubChemDeBERTa	0.9790	0.9790	0.9301	0.2657	0.9231
ChemBERT-ChEMBL	0.9790	0.9790	0.9301	0.4895	0.9301
ZINC-RoBERTa	0.9720	0.9720	0.9301	0.5210	0.9301
Text+Chem_T5-standard	0.9755	0.9755	0.9301	0.4580	0.9301
Text+Chem_T5-augm	0.9720	0.9720	0.9301	0.4825	0.9231
MolT5-base	0.9790	0.9790	0.9301	0.2727	0.9231
MolT5-large	0.9790	0.9790	0.9301	0.5245	0.9301
SciFive	0.9720	0.9720	0.9301	0.3636	0.9301
GCN	0.9510	0.9510	0.9336	0.9510	0.9510

Table 15: Accuracy metrics of different models for the Clintox task. Here, canon refers to RDKit canonicalization, hydro to Hydrogen explicit addition, kekul to Kekulization, and cycles to cycle renumbering.

Model	Orig	Cycle	Hydro	Kekul	Canon
ChemBerta	0.7571	0.7547	0.7029	0.7550	0.7280
PubChemDeBERTa	0.7602	0.7622	0.7630	0.7488	0.7690
ChemBERT-ChEMBL	0.7529	0.7511	0.7356	0.7571	0.7534
ZINC-RoBERTa	0.7591	0.7589	0.7001	0.7646	0.7560
Text+Chem_T5-standard	0.5444	0.5429	0.5203	0.5385	0.5361
Text+Chem_T5-augm	0.5434	0.5416	0.5382	0.5341	0.5387
MolT5-base	0.5190	0.5190	0.5190	0.5190	0.5190
SciFive	0.5276	0.5250	0.5224	0.5284	0.5333
GCN	0.7609	0.7609	0.7006	0.7609	0.7609

Table 16: Accuracy metrics of different models for the SIDER task. Here, canon refers to RDKit canonicalization, hydro to Hydrogen explicit addition, kekul to Kekulization, and cycles to cycle renumbering.

Model	Orig	Cycle	Hydro	Kekul	Canon
ChemBerta	0.9416	0.9417	0.9373	0.9391	0.9417
PubChemDeBERTa	0.9438	0.9431	0.9357	0.9416	0.9438
ChemBERT-ChEMBL	0.9394	0.9387	0.9145	0.9387	0.9394
ZINC-RoBERTa	0.9427	0.9431	0.9374	0.9415	0.9428
Text+Chem_T5-standard	0.9246	0.9252	0.9104	0.9213	0.9248
Text+Chem_T5-augm	0.9286	0.9282	0.9163	0.9172	0.9277
MolT5-base	0.9251	0.9231	0.9349	0.9387	0.9256
MolT5-large	0.9279	0.9206	0.8770	0.9159	0.9277
SciFive	0.9202	0.9177	0.9190	0.9262	0.9199
GCN	0.9315	0.9315	0.9245	0.9315	0.9315

Table 17: Accuracy metrics of different models for the Tox21 task. Here, canon refers to RDKit canonicalization, hydro to Hydrogen explicit addition, kekul to Kekulization, and cycles to cycle renumbering.

Model	Orig	Cycle	Hydro	Kekul	Canon
ChemBerta	0.7863	1.0001	1.5698	1.4433	0.9136
PubChemDeBERTa	0.7562	0.8178	3.1091	0.9378	0.7796
ChemBERT-ChEMBL	1.0063	1.0366	2.5114	1.1427	1.1129
ZINC-RoBERTa	0.7271	0.7581	2.2132	1.1096	0.8164
Text+Chem_T5-standard	0.9317	0.9937	7.9364	1.1054	0.8634
Text+Chem_T5-augm	0.8786	0.8531	2.1136	1.2328	0.8590
MolT5-base	1.2070	1.6411	1.4053	1.3228	1.1650
MolT5-large	1.3769	1.3078	1.8667	1.4262	1.5748
SciFive	0.7361	1.3016	2.8058	1.3166	0.9016
GCN	0.4633	0.4633	0.4633	0.4633	0.4633

Table 18: RMSE metrics of different models for the ESOL task. Here, canon refers to RDKit canonicalization, hydro to Hydrogen explicit addition, kekul to Kekulization, and cycles to cycle renumbering.

Model	Orig	Cycle	Hydro	Kekul	Canon
ChemBerta	1.7810	1.9421	6.0548	2.3067	2.0140
PubChemDeBERTa	1.3328	1.3897	4.8509	2.0143	1.7384
ChemBERT-ChEMBL	2.9604	3.0059	3.5175	3.0166	2.9847
ZINC-RoBERTa	1.2706	1.3769	7.2885	1.4714	1.3303
Text+Chem_T5-standard	1.6085	1.7771	4.4505	2.0221	1.6857
Text+Chem_T5-augm	1.4807	1.6227	3.3262	2.1552	1.9573
MolT5-base	5.0116	5.0116	5.0116	5.0116	5.0116
MolT5-large	2.2413	2.7826	4.2048	2.2003	1.7910
SciFive	1.9391	2.5557	3.9525	2.0027	2.0032
GCN	1.8461	1.8461	1.8461	1.8461	1.8461

Table 19: RMSE metrics of different models for the FreeSolv task. Here, canon refers to RDKit canonicalization, hydro to Hydrogen explicit addition, kekul to Kekulization, and cycles to cycle renumbering.

Model	Orig	Cycle	Hydro	Kekul	Canon
ChemBerta	0.6998	0.7073	1.3735	1.2675	0.8284
PubChemDeBERTa	0.6225	0.6331	0.9981	1.1520	0.7066
ChemBERT-ChEMBL	0.6087	0.6689	1.1322	1.2652	0.7180
ZINC-RoBERTa	0.5812	0.5790	1.6939	1.0694	0.6618
Text+Chem_T5-standard	0.6504	0.7083	1.2587	1.2635	0.7318
Text+Chem_T5-augm	0.6344	0.6554	1.2035	1.1295	0.7168
MolT5-base	0.6930	0.7929	1.2359	1.2024	0.8922
MolT5-large	0.8071	0.8487	1.0246	1.1913	0.8939
SciFive	0.7299	0.7780	1.0366	1.4406	0.8414
GCN	2.2289	2.2289	2.2289	2.2289	2.2289

Table 20: RMSE metrics of different models for the Lipophilicity task. Here, canon refers to RDKit canonicalization, hydro to Hydrogen explicit addition, kekul to Kekulization, and cycles to cycle renumbering.