NLPeople at *L+M-24* Shared Task: An Ensembled Approach for Molecule Captioning from SMILES

Shinnosuke Tanaka[1], **Carol Mak**[1], **Flaviu Cipcigan**[1], **James Barry**[1],
Mohab Elkaref[1], **Movina Moses**[2], **Vishnudev Kuruvanthodi**[1], **Geeth De Mel**[1]
IBM Research Europe[1] and IBM Research[2]
{shinnosuke.tanaka, carol.mak, flaviu.cipcigan, vishnudev.k, james.barry,
mohab.elkaref, movina.moses, vishnudev.k }@ibm.com, geeth.demel@uk.ibm.com

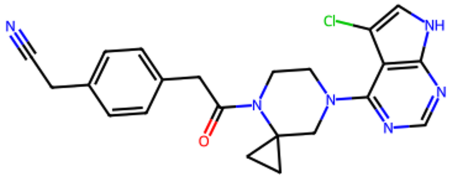
Abstract

This paper presents our approach submitted to the Language + Molecules 2024 (*L+M-24*) Shared Task in the Molecular Captioning track. The task involves generating captions that describe the properties of molecules that are provided in SMILES format. We propose a method for the task that decomposes the challenge of generating captions from SMILES into a classification problem, where we first predict the molecule's properties. The molecules whose properties can be predicted with high accuracy show high translation metric scores in the caption generation by LLMs, while others produce low scores. Then we use the predicted properties to select the captions generated by different types of LLMs, and use that prediction as the final output. Our submission achieved an overall increase score of 15.21 on the dev set and 12.30 on the evaluation set, based on translation metrics and property metrics from the baseline.

1 Introduction

Molecular design is the process of devising molecules with desired properties and functions. While this is widely practiced in fields such as drug discovery, new materials, and chemical processes, predicting the properties of designed molecules remains a challenging problem. To tackle this problem, language models trained on molecular information have gained attention (Ahmad et al., 2022). The *L+M-24* shared task (Edwards et al., 2024) involves translation between SMILES (Weininger, 1988), a string-encoded molecular format, and descriptive captions of the molecule's properties. The dataset covers four high-impact areas of molecular science: Biomedical, Human Interaction, Light and Electricity, and Agriculture and Industry, providing pairs of molecules and their corresponding captions for these properties.

An example of the data is shown in Figure 1. In this sample, specific diseases and protein properties



SMILES
N#CCc1ccc(CC(=O)N2CCN(c3ncnc4[nH]cc(Cl)c34)CC23CC3)cc1

Figure 1: A sample molecule depicted using RDKit (Landrum et al., 2024) and its caption from the training data. **Caption:** The molecule is a jak inhibitor, immunomodulator, protein tyrosine kinase inhibitor, protein kinase inhibitor and belongs to the autoimmune disease treatment class of molecules.

are described, yet the ways of describing molecular properties are highly diverse. For instance, while drug discovery seeks to generate specific information related to diseases, industrial chemistry researchers prefer to include functions of molecules such as absorption wavelengths of light. Given this variability in the desired captions, the task of generating desired captions is highly challenging.

In this paper, we describe our submission to the Molecular Captioning track. We first address the properties of SMILES as a multi-label classification problem. Predicting properties is essential for molecule captioning and offers the following advantage: a lightweight model can be built that predicts the properties of the molecules compared to fine-tuning existing large transformer-based models. Such an approach can get classification accuracy of 80% against experimental measurements with as little as 100 datapoints (McDonagh et al., 2024, 2023).

We also fine-tune $\text{LlaSMol}_{\text{Mistral}}$ and Multitask Text Chemistry T5 (Christofidellis et al., 2023) models for the end-to-end molecular captioning. We obtain the system's output by selecting the generated captions from these models based on the predicted properties. We achieve an overall increase

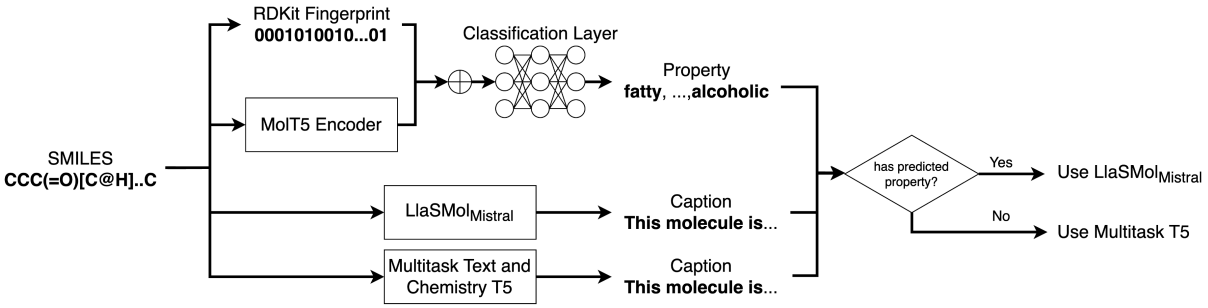


Figure 2: Overview of the submission system

score from the MolT5-Small[1] baseline of 15.21 on the dev set and 12.30 on the eval set. In the next section, we discuss some related work that inspired our contributions to this shared task.

2 Related Work

Text2Mol (Edwards et al., 2021) stands out as a pioneering study integrating modalities between text and molecules. This task involves retrieving molecules using natural language descriptions as queries. They employ the SciBERT (Beltagy et al., 2019) model to encode the text information and a Graph Convolutional Network for the molecular information. The model is based on a cross-modal attention structure and successfully integrates the two modalities.

MolT5 (Edwards et al., 2022) is a T5 (Raffel et al., 2020) based model that enables both molecule captioning and molecule generation, which generates SMILES from natural language. The model is first trained using an objective that replaces corrupted spans. This task is performed on general text data in the form of the C4 Corpus (Colossal Clean Crawled Corpus) as well as on SMILES from the ZINC-15 dataset (Sterling and Irwin, 2015). This pre-training procedure encourages the model to learn textual and chemical information. The model is then fine-tuned for molecule captioning and molecule generation using ChEBI-20 (Edwards et al., 2021), which comprises approximately 33k text-molecule pairs.

Another T5 based model, Text+Chem T5 (Christofidellis et al., 2023), aims at improving multitasking and multi-domain capabilities. This model is trained not only on SMILES and caption pairs such as ChEBI-20 but also on reaction-products pairs such as Pistachio dataset used in (Toniato et al., 2021), and experimental procedures dataset (Vaucher et al., 2019) for chemical synthesis actions. It can perform multiple tasks beyond text2molecule and molecule2text translation, including mol2mol and text2text tasks. The mol2mol tasks contain forward reaction prediction, which predicts products from given reactants, and retrosynthesis, which predicts the necessary substances for synthesis from a given chemical compound. The text2text task consists of paragraph to action, which generates sequential steps to execute a described chemical reaction. A notable aspect of this model is its ability to perform all these tasks without additional fine-tuning, using a single model instead of individual specialised models for each task. This eliminates the need to develop tailored models for each domain, achieving a unified representation of the chemical domain with one model.

3 System Description

Figure 2 shows an overview of the submission system. First, we develop a classifier to predict properties from a given SMILES string. The molecular properties are extracted using the evaluation script for the property metrics[2] by determining whether a predefined string is included in the tokenised captions using `scibert_scivocab_uncased`[3]. Based on our analysis of the extracted properties, there are 1,084 unique properties present in the training data. Since properties are extracted using string-matching, some occur together. Some co-occurances are correct biochemically, like "Biomedical disease – Heart disease" and "Biomedical disease – Diabetic heart disease". Others are not, like "Biomedical disease – Non-alcoholic

[1] https://huggingface.co/language-plus-molecules/molt5-small-smiles2caption-LPM24

[2] https://github.com/language-plus-molecules/LPM-24-Dataset/blob/main/evaluation/text_property_metrics.py

[3] https://huggingface.co/allenai/scibert_scivocab_uncased

Molecule Type	Model	BLEU-2	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
Has Predicted Props.	Multitask T5	82.15	59.49	91.64	**69.74**	60.20	87.05
	$\text{LlaSMol}_{\text{Mistral}}$	**82.66**	**59.81**	**92.27**	69.53	**60.54**	**87.70**
No Props. Predicted	Multitask T5	**43.12**	**30.58**	**52.35**	**38.51**	**50.67**	**51.87**
	$\text{LlaSMol}_{\text{Mistral}}$	35.24	24.27	48.99	35.45	47.95	45.50

Table 1: Translation metrics by molecular type on dev set.

fatty liver disease" and "Human Interaction and Organoleptics – organoleptic effect relations – fatty". This leads to chemically incorrect labelling for some molecules. A molecule whose description is "This molecule impacts non-alcoholic fatty liver disease" is not necessarily fatty in the sense of organoleptic effects, yet it is always labelled as so.

3.1 Property Classification

We build multi-label classifiers for each molecular property in the dataset. The SMILES string is converted to a binary fingerprint using the fingerprinter in RDKit 2023.9.6 (Landrum et al., 2024) with a minimum path length of 1, maximum path length of 7 and 2048 bits. SMILES strings are also provided to the encoder part of MolT5-Small, and the embedding representation is obtained by mean pooling the last hidden layer. The obtained fingerprint and embedding are concatenated and passed through a classifier consisting of three linear layers to predict the classes.

We only train the classifier on labels with over 1,500 positive examples. Because of this limitation, the *predictable subset* of the labels contains 53 properties. The classifier outputs multiple labels for each molecule that exceeds a threshold based on the Sigmoid function of the activation layer. Labels not meeting the threshold are not output; hence, some molecules may have no predicted properties.

3.2 LLMs for Caption Generation

Following the classification task, we use the SMILES string as inputs to experiment with the following methods.

Fine-Tuning LLMs We also utilise models that predict the captions directly using only the SMILES as input. In initial experiments, we found that included properties harmed performance for the Multitask T5 model[4] and as a result we did not include them. For the $\text{LlaSMol}_{\text{Mistral}}$ model, we loaded pretrained LoRA modules into the model and followed the prompt pattern in their work, which did not include properties.

Multitask T5 model is trained with a learning rate of 5e-4 and a batch size of 8 for 10 epochs on the extra training set provided by the task organiser. For fine-tuning and caption generation, we use a prompt template in Appendix A.1 Table 5, which is presented in (Christofidellis et al., 2023). $\text{LlaSMol}_{\text{Mistral}}$ is a Mistral-7b model trained on the SMolInstruct dataset by Yu et al. (2024), which covers 14 chemistry tasks including a molecular captioning task derived from the ChEBI-20 dataset. Here, the base model is frozen and additional modules are trained using LoRA (Hu et al., 2022). The LoRA component only comprises 0.58% of the full model parameters. We further fine-tuned $\text{LlaSMol}_{\text{Mistral}}$ on the *L+M-24* dataset. The prompt used is shown in Appendix A.1 Table 6. First we trained on the concatenation of train and the extra training data for 3 epochs. We then further finetuned the LoRA modules for 10 epochs on the training set.

3.3 Ensembling

We perform an ensemble by selecting generated captions from Multitask T5 and $\text{LlaSMol}_{\text{Mistral}}$ based on the result of the property classification model. If the model predicts at least one label for the target SMILES, we choose the caption from the $\text{LlaSMol}_{\text{Mistral}}$ model; otherwise, we choose from the Multitask T5 model.

4 Results and Discussion

In this section, we present the results of our classification models and generated captions using Multitask T5 and $\text{LlaSMol}_{\text{Mistral}}$.

4.1 Property Classification

When evaluated on the dev set using *only the predictable subset*, an F1 score of 97.86% was achieved. Thus, on the predictable subset, we have classifiers with a high percentage of true positives

[4] https://huggingface.co/GT4SD/multitask-text-and-chemistry-t5-base-augm

Model	Overall Increase	Translation Metric Increase	BLEU-2	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
baselines								
MolT5-Small	0.00	0.00	70.90	51.20	74.50	55.80	54.40	70.10
Meditron-7b	13.15	5.50	**79.20**	**57.60**	79.70	**60.20**	57.50	75.70
ours								
Multitask T5	**15.31**	5.23	78.22	56.73	57.28	60.17	57.28	76.27
$\text{LlaSMol}_{\text{Mistral}}$	10.59	4.68	78.84	57.17	78.82	58.79	56.50	74.87
Ensembled	15.21	**5.52**	78.70	57.04	**80.04**	60.03	**57.51**	**76.72**

Table 2: Overall increase from MolT5-Small baseline and translation metrics results on dev set.

Model	Prop. Metric Increase	Overall Prop. F1	Biomedical	Human Interaction	Agr. + Industry	Light + Electro	X-icides	Toxins	Light	Electricity
baselines										
MolT5-Small	0.00	7.88	23.33	0.56	4.36	3.27	0.00	0.00	6.54	0.00
Meditron-7b	15.70	8.93	11.94	6.51	3.04	14.22	0.00	11.05	14.10	14.34
ours										
Multitask T5	**18.67**	**19.10**	**36.97**	7.27	7.40	**24.76**	0.00	11.36	**25.26**	**24.26**
$\text{LlaSMol}_{\text{Mistral}}$	12.56	15.35	32.28	7.30	6.58	15.22	0.00	11.20	18.69	11.77
Ensembled	18.44	19.09	36.75	**7.73**	**7.65**	24.24	0.00	**12.28**	25.21	23.28
Model	**Inhibitors**	**anti-X**	**Modulators**	**Antagonists**	**Treatments**	**Agonists**	**Cancer**	**Disease**	**Combos**	
baselines										
MolT5-Small	0.09	0.00	0.00	0.00	1.70	0.00	24.27	49.94	0.00	
Meditron-7b	22.65	8.98	24.98	21.15	15.13	26.35	72.62	**82.02**	0.56	
ours										
Multitask T5	**26.04**	**10.35**	**31.11**	26.54	**19.37**	**31.71**	**73.59**	81.89	**0.93**	
$\text{LlaSMol}_{\text{Mistral}}$	14.57	5.33	15.69	12.95	9.11	19.06	70.76	81.76	0.38	
Ensembled	25.86	10.11	30.81	**26.80**	19.14	31.69	70.42	81.87	**0.93**	

Table 3: Property metric increase from MolT5-Small baseline and F1 scores of each property on dev set.

and a low percentage of false positives. When considering *all* properties in the dev set, at least one property was predicted for 69% in dev set, while no properties were predicted for the remaining 31%.

4.2 Caption Generation

Table 1 shows the translation metrics for each model, both when the classifier predicts at least one property (**Has Predicted Props.**) and when it does not (**No Props. Predicted**). When at least one property was predicted, $\text{LlaSMol}_{\text{Mistral}}$ model exceeded Multitask T5 model in 5 out of 6 metrics, excluding ROUGE-2. Conversely, when no properties were predicted, Multitask T5 significantly outperformed the $\text{LlaSMol}_{\text{Mistral}}$. Hence, based on these results, we adopted an ensemble approach where we used the captions generated by the $\text{LlaSMol}_{\text{Mistral}}$ model when at least one property was predicted, and those generated by Multitask T5 model when no properties were predicted.

Table 2 shows the overall increase, translation metric Increase and the scores of each translation metric on dev set of the two baseline models, $\text{LlaSMol}_{\text{Mistral}}$, Multitask T5 and ensembled model of $\text{LlaSMol}_{\text{Mistral}}$ and Multitask T5. Table 3 shows the property metric increase and F1 scores of each property metric. Each Increase is calculated as the average improvement from the baseline results of MolT5-Small. In the translation metrics, the Ensembled model achieved the best performance in four metrics, including the translation metric increase, indicating it has the highest performance among all models. On the other hand, in the property metrics, the Multitask T5 model showed the best performance in 13 metrics, including the property metric increase. Despite the baseline Meditron-7b model exhibiting the highest BLEU-2 score of 79.2%, our models outperformed the baseline for the property-specific F1 score. As a result, the Overall Increase was highest for the Multitask T5 model, with a score of 15.31.

Even though these predictions show higher F1 scores, the BLEU-2 score remains lower because there are numerous ways to describe molecules in natural language. This points to some features of the description which are not features of the molecule but features of the particular distribution of the dataset:

1. The order of words or phrases in a sentence, which is not essentially important, can still significantly influence these translation metric scores.

2. The `scibert_scivocab_uncased` tokeniser includes punctuation, thus mis-predicting the location of a comma or a full stop will break a correct bigram and lead to a lower BLEU score.

Team	Overall Increase	Translation Metric Increase	Prop. Metric Increase	BLEU-2	BLEU-4	Overall Prop. F1	Rank
avaliev	**27.08**	6.37	**33.99**	73.81	53.04	**26.99**	1
qizhipei	14.66	**6.45**	17.39	75.58	54.77	13.76	2
protonunfold	12.39	5.77	14.60	**75.66**	**54.98**	11.51	3
NLPeople (ours)	12.30	5.68	14.50	75.54	54.83	11.63	4
langmolecules†	10.34	5.47	11.96	75.16	54.72	9.70	8
langmolecules‡	0.00	0.00	0.00	66.82	48.29	3.23	18

Table 4: Top four results and two baseline results on the eval set. † represents the results from the baseline model, Meditron-7b, and ‡ represents the results from MolT5-Small respectively. Best results are in **Bold**, and second-best results are underlined.

3. Mis-predicting the number of properties will also reduce the BLEU score. Some of these properties are very general, such as the organoleptics, and may be correctly predicted for a molecule even if they do not exist in the ground truth caption. For example, molecules with long carbon tails will all likely taste fatty, but only the subset of those who were actually tasted by humans have the fatty caption.

Given all these features of the data, it would be interesting to create realistic performance bounds for a molecule to text model evaluated using BLEU scores, similar to the ones Crusius et al. (2024) used for regression and classification datasets by randomising over the features of the caption that *cannot* be predicted from a molecule. For example in our testing, using the *ground truth labels* in a zero-shot prompted Meditron-7b gave a BLEU-2 score of 76.36. Thus, our intuition is that we are close to saturating this benchmark, with some models achieving performance *higher* than this value.

Finally, Table 4 shows the results of the evaluation set. It includes the increases in overall, translation, and property metrics, as well as BLEU scores and property F1 scores, from the official leaderboard. Our team NLPeople's submission results from ensembling Multitask T5 and $\text{LlaSMol}_{\text{Mistral}}$. Based on the results of the property classification, out of 21,942 data points, approximately 35% used cations generated by Multitask T5, while the remaining 65% are from $\text{LlaSMol}_{\text{Mistral}}$.The team avaliev significantly outperformed other teams in the property metric, resulting in the highest overall score of 27.08. Our submission showed an increase of 12.30 overall from the MolT5-Small baseline, ranking fourth and achieving the second-highest BLEU-4 score of 54.83 among all teams.

5 Conclusion

In this work, we present our approach to the molecular captioning task. We propose combining a property classification model, LLMs for caption generation, and an ensemble method. Our results show that molecules distinguished by property classification exhibit varying strengths and weaknesses depending on the model used. This approach achieved a translation increase score of 5.52 on the dev set and 5.68 on the eval set. For property metrics, we recorded an increase score of 18.44 on the dev set and 14.50 on the eval set. The overall increase score was 15.21 on the dev set and 12.30 on the eval set, ranking 4th in this shared task.

References

Walid Ahmad, Elana Simon, Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. 2022. Chemberta-2: Towards chemical foundation models. *Preprint*, arXiv:2209.01712.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

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 *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Daniel Crusius, Flaviu Cipcigan, and Philip Biggin. 2024. Are we fitting data or noise? analysing the predictive power of commonly used datasets in drug-, materials-, and molecular-discovery.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. 2022. 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, Qingyun Wang, Lawrence Zhao, and Heng Ji. 2024. L+M-24: Building a dataset for language + molecules @ acl 2024. *Preprint*, arXiv:2403.00791.

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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Greg Landrum, Paolo Tosco, Brian Kelley, Ric, David Cosgrove, sriniker, Riccardo Vianello, gedeck, NadineSchneider, Gareth Jones, Eisuke Kawashima, Dan Nealschneider, Andrew Dalke, Brian Cole, Matt Swain, Samo Turk, Aleksandr Savelev, Alain Vaucher, Maciej Wójcikowski, Ichiru Take, Vincent F. Scalfani, Daniel Probst, Kazuya Ujihara, Rachel Walker, guillaume godin, Axel Pahl, Juuso Lehtivarjo, Francois Berenger, strets123, and jasondbiggs. 2024. rdkit/rdkit: 2023_09_6 (q3 2023) release.

James L. McDonagh, Benjamin H. Wunsch, Stamatia Zavitsanou, Alexander Harrison, Bruce Elmegreen, Stacey Gifford, Theodore van Kessel, and Flaviu Cipcigan. 2023. Machine guided discovery of novel carbon capture solvents. *Preprint*, arXiv:2303.14223.

James L. McDonagh, Stamatia Zavitsanou, Alexander Harrison, Dimitry Zubarev, Theordore van Kessel, Benjamin H. Wunsch, and Flaviu Cipcigan. 2024. Chemical space analysis and property prediction for carbon capture solvent molecules. *Digital Discovery*, 3(3):528–543.

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. *J. Mach. Learn. Res.*, 21(1).

T. Sterling and J. J. Irwin. 2015. Zinc 15 – ligand discovery for everyone. *Journal of Chemical Information and Modeling*, 55:2324–2337.

Alessandra Toniato, Philippe Schwaller, Antonio Cardinale, Joppe Geluykens, and Teodoro Laino. 2021. Unassisted noise reduction of chemical reaction data sets. *Preprint*, arXiv:2102.01399.

Alain Vaucher, Federico Zipoli, Joppe Geluykens, Vishnu Nair, Philippe Schwaller, and Teodoro Laino. 2019. Automated extraction of chemical synthesis actions from experimental procedures.

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.

Botao Yu, Frazier N. Baker, Ziqi Chen, Xia Ning, and Huan Sun. 2024. Llasmol: Advancing large language models for chemistry with a large-scale, comprehensive, high-quality instruction tuning dataset. *arXiv preprint arXiv:2402.09391*.

A Appendix

A.1 Prompt Templates

We present the prompt templates in Tables 5 and 6.

```
Caption the following SMILES: {SMILES}
```

Table 5: Multitask Text+Chem T5 Prompt with Molecule

```
Query: Describe this molecule: <SMILES> {SMILES}
</SMILES>

Response: The molecule is an imidazole derivative
with short-acting sedative, hypnotic, and general
anesthetic properties.  Etomidate appears to
have gamma-aminobutyric acid (GABA) like effects,
mediated through GABA-A receptor.  The action
enhances the inhibitory effect of GABA on
the central nervous system by causing chloride
channel opening events which leads to membrane
hyperpolarization.
```

Table 6: Prompt with SMILES and Caption for the Mistral-7b LlaSMol model.