Fine-Tuning T5 for Reaction Prediction:
Influence of Model Size and Decoding Strategy

Cody Aldaz
Stanford XCS224U
codyaldaz@gmail.com

Abstract

Chemical problems are increasingly being solved using Transformers, however the unique vocabulary, syntax, and relatively small datasets in Chemistry makes training Transformers for Chemistry a difficult task. Herein, I fine-tune a pre-trained model known as molT5 on reaction prediction, and experiment with the model size, and decoder procedure. I find that despite the general trend in the literature that pre-trained models can increase performance both molT5-base and molT5-small perform worse than a Transformer trained from scratch. The reason for this reduction in performance is currently unknown and further investigations are necessary to understand what architectural or training parameters are responsible. I also find that top-p nucleus sampling produces better top-k prediction scores than beam search. Top-p sampling was investigated to allow more diverse outputs however, the importance of top-p sampling on top-1 prediction scores was unexpected for chemical reaction prediction. Overall, these results reveal interesting details which can help us understand Transformers better and improve their performance.

1 Introduction

In recent years, researchers have developed novel applications and achieved state of the art performance in chemistry problems such as reaction prediction using the Transformer architecture.[1] However, despite the usually good performance of these models, most of the previous work trained Transformers from scratch rather than use pre-trained models like T5. This contrasts with how transformers are normally trained, since in most domains fine-tuning a pre-trained model is much more data and compute efficient (because the model already knows the grammar and syntax) than training a Transformer from scratch. The previous works also did not explore the dependence of the results on model size, or natural language generation procedure.

Therefore, for this class project I fine-tune a pre-trained model for chemical reaction prediction,[2] and I investigate the dependence of the results on model size and natural language generation procedure. Specifically, I fine-tune the pre-trained model molT5 using two model sizes, base and small, and evaluate the result with beam search and top-p (nucleus) sampling. molT5 was chosen because it is pre-trained using the highly successful T5 architecture and masked language model, and has shown promising performance in chemistry applications.[2,3] Top-p sampling is an interesting decoding strategy because it can produce more diverse but still meaningful outputs. Top-p sampling might therefore generate outputs that overall have a greater chance of correctly identifying the outcome within the number of specified return sequences.

Interestingly, molT5 performs worse than the Molecular Transformer. MolT5-base achieved 74.7% top-1 prediction score, and 78.2% top-5 prediction score using a beam search decoder. In contrast, the Molecular transformer achieves 87.6% top-1 and 92.4% top-5 prediction scores. A potential source of the difference is the Transformer size, the Molecular Transformer is much smaller than molT5-small and molT5-base (20M parameters vs 77M and 247M parameters). It may be beneficial to have a smaller transformer because the dataset is relatively small. Future work will continue to explore the hyperparameter optimization space to investigate this difference.

The rest of the paper is organized as follows: In the related works section I detail some of the exciting work that is being done with Transformers in chemistry. In the Data section I outline the datasets that I am using, and other datasets that can be used for these problems. In the Model section I

describe the molT5 model and how it was pre-trained and fine-tuned using Pytorch Lightning and the HuggingFace Library. In the experiment section I detail the experiments performed with the natural language generation procedure and model size. Finally, in the analysis section I detail the main metrics that this paper focuses on, which are BLEU, and top-K rank prediction scores, and analyze how these values change with the procedure.

2 Related Works

The Molecular Transformer was one of the first works that applied transformers to chemistry problems. In this work, the authors applied the Transformers to the chemical reaction prediction problem. Specifically, the model is trained to predict chemical products given reactants. The chemicals (i.e., molecules) are represented using the simplified molecule line input system (SMILES) which produces character strings representing the graph-like connectivity of atoms and bonds. For example, the string "CCO" represents the molecule ethanol. Remarkably, the Molecular Transformer can accurately predict the product SMILES given reactant SMILES.

The Molecular Transformer was followed up by work including papers which explored data augmentation and papers that explored transfer learning in a smaller set of reaction space.[4] Furthermore, Transformers have also been applied to the prediction of reactions to synthesize a molecule (known by chemists as retrosynthesis since it is the opposite of regular synthesis),[5] translation of experimental protocols written in English prose into discrete actions,[6] prediction of recipes given chemical ingredients,[7] molecule captioning,[2] molecule natural language generation,[2] and spectroscopic predictions.[8]

There are also relationships of these problems to other problems in natural language processing. For example, recipe generation is highly behaviorally related to culinary recipe generation, for which Transformers have also recently been applied.[9] Given the close behavioral similarity of these tasks, there may be opportunities for behavioral fine-tuning, or transfer learning, which could be improve these systems and making them more robust. The relationship between models is important reason to use a model sharing service like the HuggingFace Hub.

Lastly, some recent work has also considered efficient tokenization schemes for SMILES.[10] In most prior work that deals with SMILES, an atom-wise tokenization scheme was used, including the model system investigated herein (molT5, see Model section). However, atom-wise tokenization ignores long range relationships of molecules (e.g., when describing a cyclic molecule, the first token of the ring and last token have a strong relationship) which can more easily create erroneous output. Therefore, Smiles Pair Encoding tokenization scheme was developed, which creates more chemically meaningful tokens that can improve downstream prediction tasks.[10] It would be worthwhile to investigate this tokenization strategy in future works.

Overall, the field of Transformer learning in chemistry is highly interesting and useful, but there remain many open questions. A strong foundation of how to train Transformers effectively, and how to share these models with a broader community is an important problem that this work seeks to address.

3 Data

The amount and availability of data is one of the most difficult aspects of machine learning in Chemistry. The dataset that is used herein is the United States Patent and Trademark Office parsed by MIT researchers.[11] This is a popular and open-source database of chemical reactions used in many publications. The dataset that we use herein is composed of ~480k reactions which are split up into 409035 train, 30000 validation, and 40000 test dataset split.

Owing to the relative dearth of chemical datasets, a few other datasets are worth mentioning. The Pistachio database licensed by NextMove is another good source of chemical reactions but is proprietary, it includes a much larger and more carefully parsed chemical reactions database of reactions. The Molecular Transformer trained on this database can therefore obtain much better results. Many other proprietary datasets of chemical reactions also exists, such as Reaxys.[12] Many researchers have called for open access to chemical reaction databases.[13,14]

NextMove, also produces and licenses a database of chemical actions that they have generated from experimental procedures. This is known as the paragraph-to-actions database.[6] The database was also augmented with 1764 hand

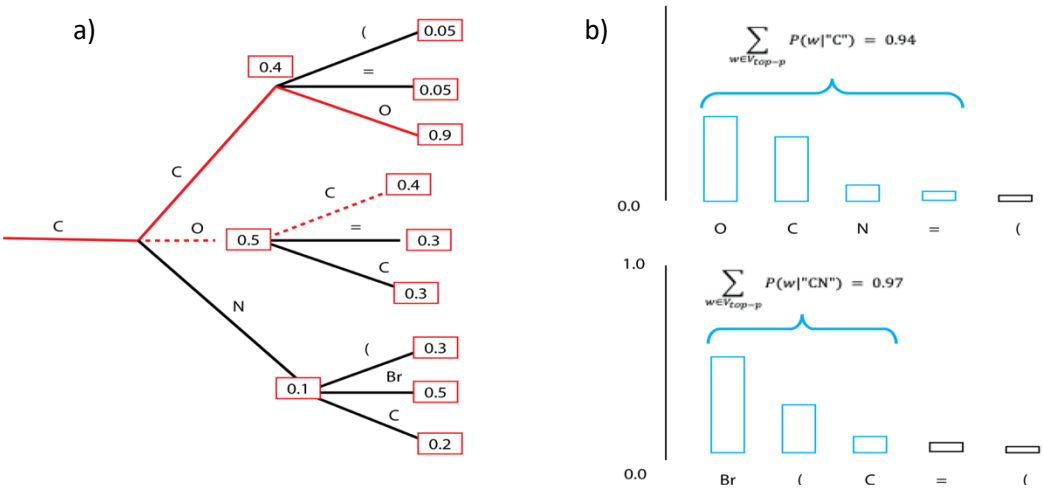


Figure 1: Illustration of decoding strategies a) beam search with number of beams=2, the full red line is the top prediction, the dashed red-line is the second prediction, the joint probability of the top-prediction is greater than the second. The black lines are not followed. b) Top-p sampling chooses from the smallest possible set of words whose cumulative probability exceeds the probability p. The probability mass is redistributed among this set of words.

annotated samples.[6] A known issue of this database, however, is that it lacks sequences which refer to distant previous steps like "Prepare Vessel A, Prepare Vessel B, Add Vessel A to Vessel B". These are common in chemical reactions. A potential good source of clean, precise experimental procedures would be the Organic Syntheses journal since all procedures are very carefully written, verified and follow a strict format.

The paragraph-to-action database has also been modified to create a database of reactant smiles-to-actions.[7] As mentioned in the previous section, the culinary recipe dataset used for recipe natural language generation has interesting overlap with this task and is worth exploring as well.[9] Another database has recently been developed for molecule captioning and molecule generation.[2] Unfortunately, this database only contains 33,010 molecule-description training examples. Which likely explains relatively poor performance of the model.

4 Model

Herein, I fine tune the molT5-base and molT5-small Transformers using the HuggingFace Libraries and Pytorch Lightning. MolT5 was downloaded from the HuggingFace Hub (https://huggingface.co/laituan245/molt5-base). MolT5-base and MolT5-small have 247 M, and 77 M trainable parameters respectively. MolT5 was previously trained from scratch using a masked language model on mixed corpus of SMILES and English prose (the C4 dataset). MolT5 uses a combination of atom-wise and sentence-piece tokenization. Because of MolT5 mixed corpus training it is highly suitable for many types of chemistry tasks such as reaction prediction, recipe generation, and molecule captioning. However, additional tokens would need to be added to appropriately train on numeric inputs. Some of the training hyperparameters used to train molT5 are as follows:

max_seq_length=512,
learning_rate=3e-4,
weight_decay=0.05,
adam_epsilon=1e-8,
warmup_steps=16000,
train_batch_size=16,
gradient_accumulation_steps=2,

and early stopping with min change=0.001 and patience = 3. The validation metric used average loss. The model was trained using a Nvidia 3080-TI. All code is available at https://github.com/craldaz/T5-Reaction-Prediction.

5 Experiments

The molT5 Transformer is fine-tuned for reaction prediction using two model size, base and small, and is evaluated with two beam search decoder strategies (Figure 1). The first strategy is a standard beam search with number of beams=10, and 5 return sequences, which was also utilized in the original Molecular Transformer paper[1] and the second strategy is top-p (nucleus) sampling, with

p=0.95, k=20, and 5 return sequences. Top-p sampling can produce more diverse output because its samples from the most probable next words rather than follow a strict next best outcome. This introduces randomness into the result but can overcome repetitive and similar outputs produced by beam search.

6 Analysis

The results of the experiments are presented in Table 1, and training curves are provided in the Appendix. As evident by the avg_val_loss continuing to decrease neither model is overfit. Notably, the performance of molT5 was worse for all experiments than the reference calculations. The difference in prediction quality between molT5-Base and molT5-small is relatively small, most likely because the dataset is proportionally small, so the increase in the number of parameters is not that helpful, or perhaps even harmful. MolT5-Base was about 3-4x more expensive to train owing to its ~3x increase in size.

The experiments also reveal that top-p sampling improves the prediction quality. This is surprising because top-p sampling was not expected to improve the performance of top-1 rank, but rather increase overall diversity in the top-5. The best result should not be improved by random selection of suboptimal choices. This indicates that the beam search is too greedy, and perhaps a larger beam search will improve results.

Table 1

Model	BLEU[1]	top-1 (%)	top-2 (%)	top-3 (%)	top-5 (%)
Reference[2]	-	87.6	90.6	91.5	92.4
Base (BS)[3]	0.93	74.7	78.7	79.9	80.9
Small (BS)[3]	0.92	71.7	75.8	77.1	78.2
Base (top-p)	0.96	82.5	85.8	87.2	88.6
Small (top-p)	0.95	82.0	85.4	86.9	88.4

[1] BLEU is evaluated for the top-1 prediction vs target
[2] Reference 1
[3] Beam Search

7 Conclusions

In this project I explored T5 for reaction prediction and experimented with the decoder. Unfortunately, molT5 performed worse than the Transformer trained from scratch. This may be due to several factors including model architecture and training procedure and further investigation into which hyperparameters are important for fine-tuning T5 for chemistry would be worth pursuing in the future. The experiments with the decoder showed that top-p sampling performed better than beam search. I speculate that this may be due to the width of the beam search.

Although my model did not achieve top performance, I learned many lessons on how to properly train a Transformer for sequence-to-sequence tasks (most of which was too trivial to be recorded here). Future work can build upon the training and evaluation examples to further improve reaction prediction and other chemistry problems.

8 References

(1) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Cent. Sci.* **2019**, *5* (9), 1572–1583. https://doi.org/10.1021/acscentsci.9b00576.

(2) Edwards, C.; Lai, T.; Ros, K.; Honke, G.; Ji, H. Translation between Molecules and Natural Language. arXiv April 26, 2022.

(3) Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P. J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arXiv July 28, 2020.

(4) Schwaller, P.; Vaucher, A. C.; Laino, T.; Reymond, J.-L. Data Augmentation Strategies to Improve Reaction Yield Predictions and Estimate Uncertainty. 6.

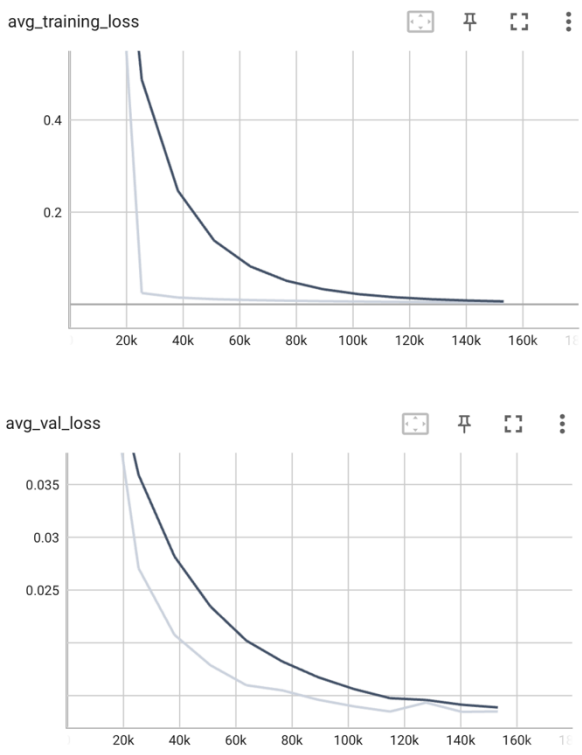
(5) Lee, A. A.; Yang, Q.; Sresht, V.; Bolgar, P.; Hou, X.; Klug-Mcleod, J. L.; Butler, C. R. Molecular Transformer Unifies Reaction Prediction and Retrosynthesis across Pharma Chemical Space. *Chem. Commun.* **2019**, *55* (81), 12152–12155. https://doi.org/10.1039/c9cc05122h.

(6) Vaucher, A. C.; Zipoli, F.; Geluykens, J.; Nair, V. H.; Schwaller, P.; Laino, T. Automated Extraction of Chemical Synthesis Actions from Experimental Procedures. *Nat. Commun.* **2020**, *11* (1), 3601. https://doi.org/10.1038/s41467-020-17266-6.

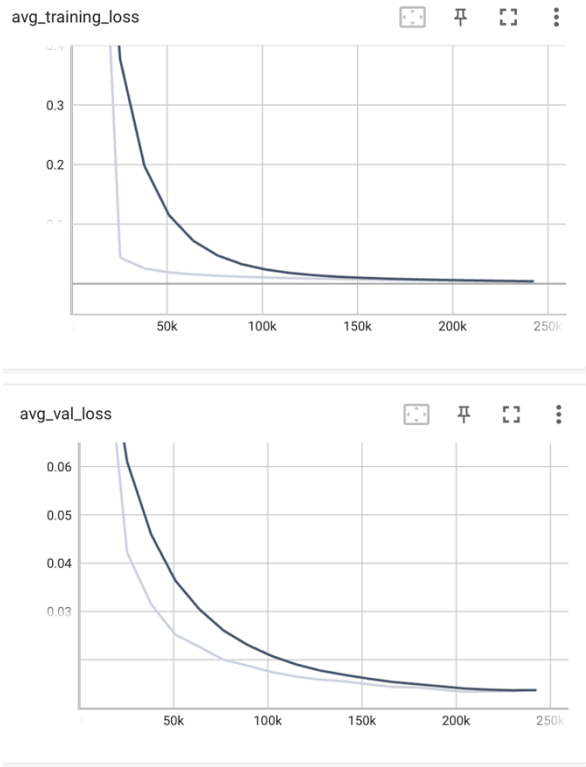
(7) Vaucher, A. C.; Schwaller, P.; Geluykens, J.; Nair, V. H.; Iuliano, A.; Laino, T. Inferring Experimental Procedures from Text-Based Representations of Chemical Reactions. *Nat. Commun.* **2021**, *12* (1),

2573. https://doi.org/10.1038/s41467-021-22951-1.
(8) Shrivastava, A. D.; Swainston, N.; Samanta, S.; Roberts, I.; Wright Muelas, M.; Kell, D. B. MassGenie: A Transformer-Based Deep Learning Method for Identifying Small Molecules from Their Mass Spectra. *Biomolecules* **2021**, *11* (12), 1793. https://doi.org/10.3390/biom11121793.
(9) Bień, M.; Gilski, M.; Maciejewska, M.; Taisner, W.; Wisniewski, D.; Lawrynowicz, A. RecipeNLG: A Cooking Recipes Dataset for Semi-Structured Text Generation. 7.
(10) Li, X.; Fourches, D. SMILES Pair Encoding: A Data-Driven Substructure Tokenization Algorithm for Deep Learning. *J. Chem. Inf. Model.* **2021**, *61* (4), 1560–1569. https://doi.org/10.1021/acs.jcim.0c01127.
(11) Jin, W.; Coley, C.; Barzilay, R.; Jaakkola, T. Predicting Organic Reaction Outcomes with Weisfeiler-Lehman Network. 10.
(12) Lawson, A. J.; Swienty-Busch, J.; Géoui, T.; Evans, D. The Making of Reaxys—Towards Unobstructed Access to Relevant Chemistry Information. In *ACS Symposium Series*; McEwen, L. R., Buntrock, R. E., Eds.; American Chemical Society: Washington, DC, 2014; Vol. 1164, pp 127–148. https://doi.org/10.1021/bk-2014-1164.ch008.
(13) Baldi, P. Call for a Public Open Database of All Chemical Reactions. *J. Chem. Inf. Model.* **2022**, *62* (9), 2011–2014. https://doi.org/10.1021/acs.jcim.1c01140.
(14) Kearnes, S. M.; Maser, M. R.; Wleklinski, M.; Kast, A.; Doyle, A. G.; Dreher, S. D.; Hawkins, J. M.; Jensen, K. F.; Coley, C. W. The Open Reaction Database. *J. Am. Chem. Soc.* **2021**, *143* (45), 18820–18826. https://doi.org/10.1021/jacs.1c09820.

9 Appendix



Supporting Information 2 Training curves for molT5-base



Supporting Information 1 Training curves for molT5-small.